The Ogburn Vision Fifty Years Later

NEIL J. SMELSER

The occasion for the symposium on which this volume is based was to note trends in knowledge in the behavioral and social sciences since the publication in 1933 of *Recent Social Trends in the United States*. That massive book was the report of a special committee of social scientists commissioned in 1929 by President Herbert Hoover to conduct a survey on the subject. It was a monumental undertaking, the last in a series of efforts of the Hoover administration to augment the knowledge base for social policy. My assignment is to try to capture the main vision of the report and to indicate the ways in which that vision has changed in the half-century since its publication.

President Hoover's own account of the reasons for deciding to launch the commission is terse. He spoke of the requests of "a number of interested agencies" (Myers, 1934:193), and he said that "the country [in 1929] was in need of more action in the social field." He added, however, that "our first need was a competent survey of the facts in the social field." Then, upon its completion he described it as "the first thorough statement of social facts ever presented as a guide to public policy," adding, however, that "the loss of the election prevented me, as President, from offering a program of practical action based upon the facts" (Hoover, 1952:312).

Hoover's account reveals his engineering view of social life: first the facts, then application based upon the facts. Later I will show how closely this mentality corresponded to that of the Ogburn committee itself.

ities. Its capacity to generate gate receipts and its value as an advertising medium are assets that cannot be ignored.

In his chapter on "Education," Charles H. Judd quoted with approval Henry Pritchett's condemnation of the consequences of competition in sports (p. 377):

> Every college or university longs for a winning team. . . . The coach is on the alert to bring the most promising athletes . . . to his college team. A system of recruiting and subsidizing has grown up. . . . The system is demoralizing and corrupt . . . the stric organization and the tendency to commercialize the sport have taken the joy out of th game.

Second, and in like spirit, there were many other statements that als might have been written today, even though we know how much thing have changed in 50 years. In one of the chapters, entitled "The Activitie of Women Outside the Home," S. P. Breckinridge concluded that "women's role in the American community has undergone redefinition during th past thirty years" (p. 709). She mentioned industrial advances, the rise c specialized services, and the decreased size of the family as having eliminated many of women's household activities. As a result, she noted th "large numbers of women through necessity or choice are seeking a ne place in the economic system." Moreover,

> the shift is not being made without revolutionary changes in attitudes with regard women's responsibilities under the changed surroundings of their lives. Their ne position . . . is giving women a share in the entire life of the community.

Third, and with the aid of historical hindsight, the reader cannot fail

[1] The report was identified with the name of Ogburn even at the time of its publication (Duffu 1933).

notice some obviously slighted topics. The committee acknowledged that the Great Depression of the time "is not explained," though apprehensive mention of its ravages appears from time to time. A generous interpretation of this is that the Great Depression struck only a few months before the committee was formed, and that the committee was as confused as the rest of the nation by the tragedy. Also, many ideas (Keynes's theory of unemployment) and measuring techniques (national economic accounts), helpful in understanding depressions, were not yet invented. In addition, however, the Depression was the largest political issue of the day, and Ogburn was insistent on presenting facts neutrally and avoiding politically sensitive issues, whether by temperament or out of deference to the President.[2]

The same reason might account for the virtual absence of materials on race and ethnic relations—though one chapter dealt with racial conditions—which seems surprising in light of the presence on the committee of Howard Odum, the day's leading sociologist of the South. It is inconceivable that such a report could be written today without major attention devoted to racial and ethnic issues. In addition to the possibility that race and other controversial areas were soft-pedaled, it should be remembered that race relations were then still largely regional rather than national, that the political mobilization of blacks was in its infancy, and that neither politicians nor social scientists had begun seriously to challenge the racist foundations of American social life—all of which would contribute to the low visibility of racial problems.

THE OGBURN VISION OF SOCIAL PROCESS

One reviewer of *Recent Social Trends* remarked that "the Committee findings are so unified and eloquent as to give the impression of single authorship" (Mallery, 1933:211). That authorship was largely Ogburn's. It is remarkable to observe the degree to which he dominated the committee report. Its main statement echoes his perspectives and theories published earlier and later, and the chapters by others frequently echo those perspectives and theories. It is generally fair, therefore, to treat the report as manifesting the Ogburn vision of the social sciences.

How best to characterize this vision? It is a view that begins with the identification of social anomalies and problems that arise through irregular

[2] On this subject, and on Ogburn's conflicts with fellow committee members Wesley Mitchell and Charles Merriam on the question of the independence of the committee from presidential involvement, see Harold Orlans (1982) and Barry D. Karl (1969, 1974). Among the chapter authors, Robert Lynd broke most conspicuously from Ogburn by insisting on stressing normative and political issues.

continuity jective facts change
and lags)

Each ingredient leads to the next, and thus constitutes a more or les
articulate theory of change. In the remainder of my remarks I intend
take up each ingredient (as well as the transitions between the ingredient
and present a capsule statement of the committee's view, then indicate ho
that view has altered over the decades, mainly as the result of ongoi
social science research and theory development.

SOCIAL CHANGE

One of Ogburn's most notable contributions as a social scientist is t
notion of "cultural lag," which enjoyed great influence in the social s
ences for a long period and is still important in the literature on soc
change (Ogburn, 1922). The kernel of this theory finds expression early
the report itself (p. xiii):

> Not all parts of our organization are changing at the same speed or at the same tin
> Some are rapidly moving forward while others are lagging. These unequal rates
> change in economic life, in government, in education, in science, and religion, m
> zones of danger and points of tension.

More particularly, Ogburn saw changes in technology as well as econon
and governmental organization leading the way of change in modern tim
with the family and church having declined in social significance.

The image of society evoked by this notion is what sociologists call "
functionalist view," namely, that the different parts of social organizat
stand in systematic—whether harmonious or disharmonious—relations
to one another, and that changes in one call for changes in another. T
view of society, in various forms, dominated a number of the social scien
for several decades and still represents a major theoretical position. S
sequent research and theory development, however, have demonstrate
to be both overdrawn and incomplete. Comparative research on the re
tionships between economy and family, for example, have demonstra
that even in the face of very rapid industrialization, some traditional fam
forms, far from being "zones of danger and points of tension," persist

anisms. The Japanese family is the classic case in point. The implication of this kind of research is that the notion of "fit" among the various parts of society is weaker than the functionalist view would imply, and that many more diverse combinations of structures are possible. A second line of criticism and reformulation runs as follows: It is not so much the "fit" or "misfit" between different structures that account for pressures for persistence and change as it is the power positions of groups or classes with vested interests and the outcomes of political struggles among these groups. This second line of development is seen as exposing and correcting for the political naiveté, if not conservatism, of the functionalist position.

SOCIAL PROBLEMS

According to the Ogburn vision, social problems emerge as manifestations of objective social situations—i.e., discontinuities and lags. For example, the automobile, a material advance, generated an outward drift of the population into suburban areas; the consequent problem was that the central districts were "left to the weaker economic elements and sometimes to criminal groups with resultant unsatisfactory social conditions" (President's Research Committee on Social Trends, 1933:xlii). In another example, the committee attributed increasing divorce rates to the fact that the family had fewer economic and other functions, which weakened personal ties among its members.

In the ensuing decades social scientists have become more sophisticated in their understanding of what constitutes a social problem. We now see that social problems emerge as a complex process of interaction between "objective" social conditions, the criteria people bring to bear in evaluating those conditions, and the success or failure of efforts of interest groups to push their particular criteria forward. Consider another example from the report. In their chapter on "The Population of the Nation," Thompson and Whelpton brought up the topic of the quality of the population. They argued that the differential birthrate among the social classes had resulted in "some deterioration in the biological soundness of the national stock" (a social problem). Their position on this matter was simply that "as soon as any agreement can be reached about the method by which 'undesirables' can be selected from the population, they should be prevented from propagating" (President's Research Committee on Social Trends, 1933:56). We would now regard this view as hopelessly naive. The quality of the population is not some kind of objectively given problem. It is a problem for some (eugenicists) and not a problem for others (the right-to-life movement) because the ideological priorities of the two groups—in the name of which

That kind of consensus rarely exists. We now know that social problems are not matters of objective fact but matters of an uncertain, disputed set of both facts and principles. Recognizing this, we can appreciate why such a large proportion of the debates about social problems are debates *not* about the existence of facts but about symbols, about the legitimacy of the competing sets of criteria by which a factual situation will or will not qualify as a genuine social problem.

DOCUMENTATION BY OBJECTIVE FACTS

In his introduction to *Recent Social Trends*, Herbert Hoover spoke of his desire "to have a complete, impartial examination of the facts" in the report. In a way this phrase encapsulates the mentality of the social sciences in the early twentieth century—the acme of positive science, which regarded empirical facts as objective things, waiting to be observed, recorded, and quantified. This mentality manifested itself in a variety of different ways. To name a few:

- the pioneering efforts to develop measures in psychology and education, including the work of Thurstone on measurement of attitudes and Terman on the measurement of intelligence.
- the reaction of the institutional economists (among them Veblen and Commons) against what they regarded as the abstract, disembodied theory of classical economics; as part of this polemic they insisted on the empirical study of economic life in concrete institutions.
- in anthropology the reaction of the diffusionists (especially Boas) against classical evolutionary theory, and their insistence on detailed, empirical studies of the movement of cultural items and artifacts from culture to culture.
- Ogburn's own dismissal of classical evolutionary theory as speculative and wrong,[3] and his insistence that the study of evolution must rest on the

[3] Ogburn wrote that the theory of "the inevitable series of stages in the development of social institutions has not only not been proven but has been disproven" (Ogburn, 1922:57).

"actual facts of early evolution" (Ogburn, 1922:66). Ogburn (1929) celebrated the rise of scientific social science in his presidential address to the American Sociological Society in 1929, stressing its emphasis on objective measurement, verification and truth, and its separation from methods in other areas such as ethics, religion, education, and propaganda.

Not everybody found comfort in this position. Pitirim Sorokin, sociologist at Harvard, in a savage review of *Recent Social Trends* in 1933, bemoaned what he called "holy and immaculate quantification":

> In the future some thoughtful investigator will probably write a very illuminating study about these "quantitative obsessions" of a great many social scientists, psychologists, and educators of the first third of the twentieth century, tell how such a belief became a vogue, how social investigators tried to "measure" everything; how thousands of papers and research bulletins were filled with tables, figures and coefficients; and how thousands of persons never intended for scientific investigation found in measurement and computation a substitute for real thought. . . .[4]

Be that as it may, Ogburn's preference for stressing objective facts, apart from opinions and value judgments, held sway in the report itself. The chapters and monographs, the committee said, "present records, not opinions; such substantial stuff as may serve as a basis for social action, rather than recommendations as to the form which action should take" (President's Research Committee on Social Trends, 1933:xciv). The contributors, moreover, were "bound strictly by the limitations of scientific methods," and if they occasionally strayed beyond these limitations the reader could see clearly when they were giving their own opinions (p. xcv).[5]

Even at the time, this "factual-statistical" representation of the world was regarded by others besides Sorokin as wanting. Adolph Berle, a member of Franklin D. Roosevelt's brain trust, commented that the report "has the barrenness of . . . statistical measurement . . . the desire for objectivity has been carried entirely too far." And Charles Beard, the historian, remarked that "the results [of this report] . . . reflect the coming crisis in the empirical method to which American social science has long been in bondage" (Orlans, 1982:9). And in the decades since the acme of Ogburnian positivism we have come to view the world of empirical facts not so much

[4] Throughout his review Sorokin assaulted the Ogburn committee report for its multiplication of meaningless quantitative tables and citations. In a rejoinder Ogburn countered with the assertion that "only one-tenth of the space is taken up with tables," a statement that constitutes a kind of ironic confirmation of Sorokin's plaint.

[5] Ogburn wrote a short methodological "note" on the necessity to separate facts and opinions sharply from one another, but this was not published as part of *Recent Social Trends*, probably because not all of the members of the committee subscribed to his position (Bulmer, 1983).

moreover, has resulted from developments both at the level of theor of empirical research. At the theoretical level, early critics of positi such as Talcott Parsons (1937), argued that facts could not be viewed from the conceptual framework by which they are evoked. In his influ work on the history of science, Thomas Kuhn (1970) argued that scientific facts and scientific knowledge are relative to the kinds of digms invented and employed by scientists. And more recently critic Jürgen Habermas have hammered away at exposing the ideologica political foundations of "objective science." The cumulative effect of kinds of intellectual development has been to effectively erode the pos dream of the early twentieth century.

At the level of social research our assessment of "facts" has also be more sophisticated. The dominant approach, of course, is still th behavioral and social sciences are *empirical* sciences above all, ar have improved our measurement techniques and data bases enormo But social scientists no longer conceive, as a Durkheim or an Ogburn have done, of the crime rate as a "social fact" to be observed. We k on the basis of empirical research, that a "crime rate" is a vastly dif phenomenon, depending on whether the investigator consults police re observes police in action, asks people whether they have ever been v of crimes, or whether they have ever committed crimes. We know als every one of these measures is defective in different ways.

We know that there is no such "thing" as public opinion, which c measured scientifically by randomly sampling a portion of the popu and interviewing them on a given set of issues. Research has show results of such surveys vary significantly depending on how the que are asked, what kinds of people do the asking (whites or blacks, m women, investigators dressed in suits or investigators dressed in dirty j and how people distort their responses on sensitive issues (such as much they smoke, drink, or use drugs) (Cannell and Kahn, 1968). We also come to acknowledge that certain ideological assumptions or are built into some of the measures we use. For example, the fact th the sample survey, we give equal weight to all respondents in ana data reflects a kind of "democratic" assumption that each person's counts as much as another's—an unrealistic assumption given wh know about actual patterns of participation, influence, and power, e democratic societies; it is the (perhaps unwitting) translation of the ele principle of a democracy into a "one-person, one-response" assum

Interestingly, these kinds of acknowledgments make simultaneous both greater humility and greater sophistication on the part of soc vestigators. We are cognizant of the many sources of measurement error that are generated in the creation and study of social data and in its assessment by investigators (Turner et al., 1984). By the same token, however, investigators are now equipped systematically to take measurement errors into account when representing and statistically manipulating data, by using techniques that would not come to mind within a simple positivistic perspective.

SOCIAL INVENTION

According to the Ogburn vision (President's Research Committee on Social Trends, 1933:lxxi) the massive accumulation and description of social facts can reveal the broad range of social problems generated in a society undergoing rapid and irregular social change. These problems, moreover, "can be solved only by further scientific discoveries and practical inventions."

The imagery of a scientific invention—as well as its application—pervades the Ogburn vision of social reform and the amelioration of social problems. In the chapter on "The Influence of Invention and Discovery," Ogburn and S. C. Gilfillan wrote that "there are social inventions as well as mechanical ones, effective in social change" (p. 162). They gave as examples the city manager plan, group insurance, installment selling, the passport, and universal suffrage.

The committee (1933:lxxiv) envisioned the need for a massive effort in the field of social invention:

> If one considers the enormous mass of detailed work required to achieve the recent decline in American death rates, or to make aviation possible, or to increase per capita production in farming, one realizes that the job of solving the social problems here outlined is a job for cumulative thinking by many minds over years to come. Discovery and invention are themselves social processes made up of countless individual achievements.

Read today, this link between knowledge about social problems and social invention appears somewhat mechanical and politically naive. First, little attention is given to the exact mechanism that provides the transition between the accumulation of knowledge and social invention. In his presidential address to the American Sociological Society in 1929, Ogburn (1929:5–6) outlined a simple model. Science, he said, is an accumulation of thousands of verified "bits and pieces of new knowledge." He envisioned that this would occur through careful, patient, and methodical work, much of which could and would be carried out by "dull and uninteresting persons." Once in a while, "one of these little pieces of new knowledge great significance" of an empirical finding derives from the fact that it demands a substantial change in the way we formulate our general understanding of the world—in short, in the way we formulate theory. Typically a "discovery" is the verification of findings that cannot be accommodated by an accepted scientific framework. Or, alternatively, a "discovery" involves a reformulation at a theoretical level, such that heretofore unrelated empirical findings can be related to one another and explained within a new framework or by a new principle. Put another way, scientific discovery always involves a *relation* between empirical findings and theoretical formulation, not an accumulation of empirical findings (Kuhn, 1970).[6]

Furthermore, with respect to "social inventions" a different set of processes needs to be invoked. Consider the social invention of universal suffrage—one of Ogburn's examples. It *is* an invention in the sense that it is a contrivance designed to facilitate the operation of the democratic process. But the role of knowledge in the crystallization of such an invention is a limited one. Much of the "knowledge" involved has not been scientific in the sense of having been proven or verified; it has been more in the nature of lore associated with democratic philosophies, which takes the form of assumptions about the workings of political influence and power. Furthermore, the dynamics of the invention were not the dynamics of assembling knowledge so much as the historical struggles of different kinds of classes and groups for access to the political systems of democracies.

More generally, social inventions appear to be the invocation of established or imputed knowledge *in relation to* some desirable social goal or social value. Consider the historical "invention" of desegregated education by the United States Supreme Court in *Brown v. Board of Education* in 1954. In that decision, justices cited a wide variety of social-science findings to the effect that separate facilities engender feelings of inferiority in blacks.

[6] For an earlier statement of the relations between empirical findings and theory in the social sciences, see Robert K. Merton's two essays, "The Bearing of Sociological Theory on Empirical Research" and "The Bearing of Empirical Research on Sociological Theory" (Merton, 1968:139–171).

But as Judge David Bazelon (Eisenberg, 1969:374) argued, reliance on these findings might have misstated the true basis for the case:

> In 1896 the court had approved the "separate but equal" doctrine. While the country might then have lacked the sophisticated studies available in 1954, any honest person would have conceded at the time of *Plessy v. Ferguson* that segregation undoubtedly would have made Negroes feel inferior. The assumption of inferiority was the rationale for the practice; no black man could help but perceive that separate train cars and separate schools kept him in his place.
>
> Since we already knew what Kenneth Clark and others told us, the public could justly ask of the Supreme Court in 1954, why the law had changed. The answer, of course, was that our values had changed. *Plessy v. Ferguson* was discarded not because social scientists told us that segregation contributed to feelings of inferiority, but because by 1954 enough people in this country believed what they did not in 1896—that to thus insult and emasculate black people was wrong, and intolerable, and therefore, a denial of the equal protection of the law to blacks.

In the area of social inventions, as in other areas, the committee's insistence on the neutrality of scientific knowledge and on its separation from matters of opinion involved a cost. In this case the cost was to miss a great part of the intricate interplay between knowledge—whether imputed or established—and the political and cultural dynamics of society.

APPLICATION BY POLICY CHANGE

Toward the end of its main report, the committee (p. lxxiii) noted with approval the "increasing penetration of social technology into public welfare work, public health, education, social work and the courts." In addition, it called for the formation of groups through the Social Science Research Council to bring technical advice to decisionmakers, and perhaps the formation of a national advisory council to focus on "the basic social problems of the nation."

We have seen, in the discussion immediately preceding, that to invoke the imagery of technology in the formation of social policies is both limiting and misleading. The same can be said when that imagery is carried over to the implementation of social policies. Two observations are in order on this score.

The first has to do with the adequacy of knowledge in the name of which policies are implemented. The putative knowledge cited in the *Brown v. Board of Education* case was that integrated school facilities would lead to a decrease in feelings of inferiority on the part of blacks. Scores of studies on the self-esteem of black children in diverse settings tell us that so many contingencies affect self-esteem—class, neighborhood, the behavior of individual teachers, the fortunes of the movement to improve conditions for in the ongoing flow of social life, other things are not constant, and precis prediction of consequences is impossible because of the interaction amon multiple forces.

A second complexity arises through the fact that any kind of policy when implemented, is likely to generate a variety of unanticipated sid effects, not all of which are predictable or likely to be beneficial. Conside only one example, that of attempting to ameliorate the incidence of suicid in society. One feasible policy would be to attack intensively the socia conditions of certain high-risk groups, such as the elderly, with the aim o reducing feelings of isolation, desertion, and despair. In implementing th kind of policy, a community might embark on a program of establishin senior citizen clubs as social centers, and making individual agencies, suc as suicide prevention centers, more available to them. Integrating the elderl into more meaningful social communities might decrease the incidence o suicide. But in addition, it might facilitate the formation of more defini political groups among the elderly, which are traditionally antipathetic educational programs that call for the passing of school bonds, as well to community health programs such as the fluoridation of drinking water- to programs, that is, that represent the implementation of *other* social goal usually considered also worthy by the planners sponsoring the suicid prevention efforts. Knowledge of the diversity of consequences of differe programs may in fact result in more intelligent setting of priorities in pla ning. In any event, it provides a different and better model for plannin than that of the direct application of bits of knowledge toward the soluti of specific problems.

SOCIAL AMELIORATION

The last link in the chain of social process is the ultimate impact knowledge on society's welfare. As indicated earlier, the committee (pp. xl xliii) was apprehensive about the trend toward higher divorce rates American society; "our culture may be conducive to further increases divorce unless programs are instituted to counteract this tendency." T ning and programs is itself a contingent matter. Just as the Ogburnian vision of what constitutes a social problem rests on the committee's imagined consensus on values, so does its notion of amelioration. In areas where widespread consensus on values obtains in society—for example, the health of the population—programs like mass immunization are likely to be uncontroversial and widely regarded as ameliorative. When, however, such consensus is lacking, one group's amelioration is another group's deterioration. Even the Ogburn committee's invocation of the value of ''family stability'' as a consensual matter could be and has been challenged by those committed to communal and other arrangements believed to be superior to the traditional family. When consensus is lacking, moreover, debate comes to focus not only on the consequences of programs but on the relative legitimacy of the competing cultural values by which we judge those consequences. In this respect, the assessment of consequences is as deeply embedded in the political and cultural dynamics of a society as is the identification of social problems.

A CONCLUDING NOTE

We end with a kind of paradox. Even though the Ogburn report seeks legitimacy mainly from the framework of positive science, its vision of the social process is characterized by a number of items of faith: faith in the capacity of objective knowledge to identify social problems, faith in the capacity of cumulative knowledge to result in social inventions, and faith in the capacity of those inventions to solve the social problems. That particular set of faiths permitted the committee to be simultaneously naive and pretentious—at least as judged by our contemporary understanding—about the role of the behavioral and social sciences in social policy. The same set of faiths permitted the committee to define social and behavioral scientists as simultaneously disembodied from the political process and essential ingredients to that process. Such are the paradoxical consequences of the positivist-utilitarian view of the relations between science and society.

Today I believe we would acknowledge the tremendous importance and utility of the social sciences in the social and political life of the nation. In its first report (Adams et al., 1982), the Committee on Basic Research in the Behavioral and Social Sciences acknowledged this and pointed to three areas in particular: technical contributions in the information-generating process, such as sample surveys and standardized testing; changes in the way we do things, such as administer therapy, predict economic trends, and run organizations; and changes in the way we think about things such as poverty, race, social justice, and equity in society. Yet the present committee, mindful of the kinds of complexities and contingencies that have been touched upon in this discussion, regarded these not as utilitarian applications of bits of scientific knowledge, but rather as arising from and intertwined with the social purposes and cultural aspirations of the nation as a whole. As a result of change in our thinking about the relations between science and society, I believe we have become, paradoxically, both more sophisticated in our research design and measures and less pretentious in our aspirations than we were 50 years ago.

REFERENCES

Adams, Robert McC., Smelser, Neil J., and Treiman, Donald J., eds.

1982 *Behavioral and Social Science Research: A National Resource*. Washington, D.C.: National Academy Press.

Bulmer, Martin

1983 The methodology of early social indicator research: William Fielding Ogburn and "Recent Social Trends." *Social Indicators Research* 13 (2):109–130.

Cannell, Charles F., and Kahn, Robert L.

1968 Interviewing. Pp. 526–595 in Gardner Lindzey and Elliot Aronson, eds., *Handbook of Social Psychology*. Vol. II. Reading, Mass.: Addison-Wesley.

Duffus, R. L.

1933 Whither? A survey of the nation's course. *New York Times*. January 8.

Eisenberg, L.

1969 Judge David L. Bazelon. *American Journal of Orthopsychiatry* 39:372–376.

Hoover, Herbert

1952 *The Memoirs of Herbert Hoover, the Cabinet and the Presidency, 1920–1933*. New York: Macmillan.

Karl, Barry D.

1969 Presidential planning and social science research: Mr. Hoover's experts. *Perspectives in American History* 3:347–409.

1974 *Charles E. Merriam and the Study of Politics*. Chicago: University of Chicago Press.

Kuhn, Thomas

1970 *The Structure of Scientific Revolutions*. Chicago: University of Chicago Press.

Mallery, Otto T.

1933 Review. *The Annals of the American Academy of Political and Social Science* 166.

Merton, Robert K.

1968 The bearing of sociological theory on empirical research. The bearing of empirical research on sociological theory. Pp. 139–171 in Robert K. Merton, *Social Theory and Social Structure*. New York: Free Press.

Myers, William Starr, ed.
1934 *The State Papers and Other Public Writings of Herbert Hoover*. Garden City, N.Y.: Doubleday, Doran & Co.

Ogburn, William F.
1922 *Social Change: With Respect to Culture and Original Nature*. New York: B. W. Huebsch.

Ogburn, William F.
1929 The folkways of a scientific sociology. *Studies in Quantitative and Cultural Sociology*. Washington, D.C.: Publications of the American Sociological Society 24:1–11.

Orlans, Harold
1982 Social Scientists and the Presidency: From Wilson to Nixon. Unpublished draft dated 5/17/82.

Parsons, Talcott
1937 *The Structure of Social Action*. New York: McGraw-Hill.

President's Research Committee on Social Trends
1933 *Recent Social Trends in the United States*. New York: McGraw-Hill.

Smelser, William T., and Smelser, Neil J.
1981 Group movements, sociocultural change, and personality. Pp. 625–652 in Morris Rosenberg and Ralph H. Turner, eds., *Social Psychology: Sociological Perspectives*. New York: Basic Books.

Sorokin, Pitirim A.
1933 Recent social trends: A criticism. *The Journal of Political Economy* 41(2).

Turner, Charles F., and Martin, Elizabeth, eds.
1984 *Surveying Subjective Phenomena*. 2 vols. Panel on Survey Measurement of Subjective Phenomena, Committee on National Statistics, National Research Council. New York: Russell Sage Foundation–Basic Books.

Measuring Social Change

ALBERT J. REISS, JR.

INTRODUCTION

Surely among the most influential models of social change was that developed by William Fielding Ogburn (1922b). Ogburn described a process of invention followed by cultural change, followed by social disorganization, and finally social adjustment (Ogburn & Nimkoff, 1940:877). Ogburn concluded that public policies and interventions meant to guide modern social change would depend heavily upon the development of a unified national statistical system to collect and process information about social trends (Ogburn, 1929:958). Although Ogburn's vision of a unified statistical system has not been realized, he may well have regarded this as but a lag in adjustment to which all inventions give rise.

This essay does not attempt to assess systematically Ogburn's (1922b) theory of social change, his contributions to our understanding of social trends (1928–1935, 1942), or the development of statistical systems (Ogburn, 1919; President's Research Committee on Social Trends, 1933). But it draws heavily upon that vital heritage. Three major questions are addressed: (1) How do inventions, especially those of the behavioral and social sciences, affect social changes and adaptations? (2) How do social changes affect measurement? And, (3) How do contemporary behavioral and social science models, concepts, and methods affect our understanding of society and how it changes?

SOCIAL INVENTIONS

In Ogburn's view, inventions, particularly mechanical ones, are the source of all cultural growth and evolution. Inventions also cause disruptions in related parts of culture and in social organization, necessitating adaptations and adjustments. But these adjustments take time, and Ogburn therefore called them cultural lags, noting that "Over the long course of social evolution, measured in thousands of years, cultural lags are invisible. At any particular moment, however, they may be numerous and acute" (Ogburn in Duncan, 1964:30).

Although Ogburn emphasized that social inventions can cause social change (1934:162), his theory and his own work gave priority to mechanical inventions (1922b:76–77; Ogburn & Nimkoff, 1940:809–810).[1] This benign neglect of social inventions is coupled with Ogburn's firm conviction that the behavioral and social sciences can shorten cultural lags. Nowhere did he summarize this belief better than in his chapter on invention in *Recent Social Trends* (President's Research Committee on Social Trends, 1933:166):

> Society will hardly decide to discourage science and invention, for these have added knowledge and have brought material welfare. And as to the difficulties and problems they create, the solution would seem to lie not so much in discouraging natural science as in encouraging social science. The problem of the better adaptation of society to its large and changing material culture and the problem of lessening the delay in this adjustment are cardinal problems for social science.

Ogburn concluded an essay on trends in social science with these observations (1934:262):

> The greatest obstacles to the development of science in the social field are complexity of the factors and the distorting influence of bias. These are formidable, but certainly the trends of the present century are most encouraging, and we may look forward, because of social science, to a greater control by man of his social environment.

The relatively lesser emphasis that Ogburn placed on the role of social as compared with material technology persists to this day. Even social and behavioral scientists tend to overlook their role in processes of social change. In fact, it is quite plausible that social inventions, especially those of the behavioral and social sciences, are a major cause of change, as well as key elements in society's adaptation to change. The selective perception that limits recognition of the role of behavioral and social science inventions may indeed count as a cultural lag.

[1] Ogburn's interest in social inventions, their effects, and lags in adapting to them preceded the writing and publication of his classic study, *Social Change* (1922b). His doctoral dissertation (1912) was on child-labor legislation. While teaching at Reed College in Oregon, he became interested in the initiative and referendum as methods of direct legislation (1914, 1915). Still later, he was interested in the consequences of women's suffrage (Ogburn & Goltra, 1919). As Duncan concludes, however, this early interest in social inventions arose, in part, from political sympathies with social problems and reforms.

Underlying the major themes for this first section is a speculation th the relative contributions of the respective sciences and technologies social change are altering substantially. Modern societies have come depend heavily on the behavioral and social sciences and their technologi and cannot run without them. As material technology replaces labor, no material technology may come to dominate social change, if it has n already done so.

Major Social Inventions and Their Consequences

Ogburn was fascinated by the effect of what he distinguished as maj technological inventions such as the ship, the airplane, the internal con bustion engine, and the elevator. He also devised lists of significant soci inventions (1934:162), such as the minimum wage law, the juvenile cour Esperanto, installment selling, and group insurance. Yet he apparently nev attempted to differentiate between social and behavioral inventions wi potentially major versus those with more limited or minor effects. Son social and behavioral science inventions, nonetheless, have had such si nificant and widespread impact that one cannot imagine modern democrat societies operating without them. Two such inventions, noted in the fir report of the Committee on Basic Research in the Behavioral and Soci Sciences (Adams et al., 1982) are singled out here: human testing ar sample surveys.

Human Testing Ogburn (1950) generally attributed invention to thr fundamental causes: mental ability, social demand, and the accumulatio of cultural elements from which inventions are fashioned. To pinpoint tl origins of a particular invention is not a simple task, given the multiplici of able minds, the variation in the sources of demand, and the differe patterns that elements of the cultural base may take.

The invention of human testing is usually attributed to a nineteent century scientific interest in the study of individual differences. The histo of tests of distinctly mental abilities is better documented than other maj forms of human testing (Wigdor & Garner, 1982). Tests of mental abiliti derived from psychologists' attempts to understand differences in intel gence among individuals. Galton (1869) first devised a series of senso discrimination tests to shed light on individual differences, followed l Cattell (1890) and others who developed batteries to test sensory and mot abilities. But it was a demand within the French Ministry of Education, distinguish subnormal from normal children in Paris schools, that led Bin in collaboration with Simon (1905), to introduce the concept of mental a and scales to measure it.

Ogburn often noted that inventions diffuse more readily where there is a demand for them; the Binet-Simon scale diffused quickly. The test was translated into English by Goddard in the United States in 1908, into Italian by Ferrari in 1908, and into German by Bobetrag in 1912 (Klineberg, 1933:323). Translation was followed by revision, such as the Stanford-Binet test published by Terman and his collaborators in 1916 (Klineberg, 1933:324).

Although testing has been important to the conduct of research and was a product of psychological laboratories, its development and invention have been highly responsive to social demands arising outside the laboratory, initially by the public schools to sort children and somewhat later by the U.S. Army to screen World War I draftees. Testing is now at least as consequential for the major operating organizations in industrial societies as for the conduct of research. The testing industry is integral to four major organizational tasks: (1) *selection* of persons as employees or clients; (2) *classification* of employees or clients according to organizational tasks; (3) *assessment* of human performance within organizations; and (4) *assessment of the ''human output''* of organizations.

Ogburn distingushed primary from derivative effects of invention. Since societies and their organizations do not systematically collect and process information about such effects, even less so for social than mechanical inventions, it is far easier to identify qualitatively than to document the quantitative impact of the invention of human testing. The primary effects are clearly on employment and the management of organizations. Testing occupations generate substantial employment in the U.S. Civil Service, the Armed Forces, public and private school systems, and in large private industrial firms, most of which employ testing extensively in at least one of the four organizational tasks mentioned above, as well as in the development, production, and marketing of tests themselves.

Public controversy and litigation may surround the use of testing in organizational management. Because many organizations base selection and promotion on testing, test information can be influential in legal proceedings. The testing industry has been challenged to produce different kinds of tests as a consequence of such litigation. The courts have played a substantial role, for example, in structuring tests for selecting and promoting women and minorities in police and fire departments.

Derivative effects of behavioral and social inventions include the spur they often provide to mechanical inventions. The first high-speed printer (essential for modern computers) was developed for a scoring machine by the educational tester Lindquist. In the highly competitive educational achievement testing industry, the rapid scoring and delivery of test results invention and mechanical invention are seldom independent of one another. The design of modern control systems necessarily involves both human performance measures and technological components. The displacement of humans by computerized robots is also a replacement of some human skills by other human skills. The machine's displacement of manual or mechanical labor moves the labor force toward the cognitive skills that are most distinctively human.

It seems no exaggeration to estimate that the average person in an industrial society encounters the products of the testing industry virtually every year for the first two decades of life and in many cases for much of his or her career. Even where not subject to standardized tests, occupational life is controlled by elementary concepts of ability and achievement developed in testing. Increasingly, testing concepts enter the debate over major issues in society, such as the recent controversy over merit pay for teachers—especially whether merit can be based on testing teacher performance.

Aside from the considerable effect on every other sector of society, the invention of testing precipitated many new inventions in statistics and other behavioral and social sciences. These inventions have significantly affected the conduct of research, and the results of that research have in turn affected society. The early testing of intelligence and mental abilities led to Spearman's attention to the reliability of measures and his positing of the G factor in intelligence (Spearman, 1904); this development gave rise to factor analysis, especially with Holzinger's (1930, 1931) development of the bifactor method (through a study with K. Pearson and collaboration with Spearman, 1925). A variety of statistical factoring methods were soon invented as the concept of intelligence changed with empirical testing, including multiple-factor methods (Thurstone, 1931, 1935) and principle component methods (Kelley, 1928, 1935; and Hotelling, 1933). As factor analysis was extended to other human traits and characteristics, e.g., human emotions (Burt, 1915, 1939), attitudes, and opinions, awareness of its limitations led to statistical inventions for discerning latent structures (Guttman, 1950; Lazarsfeld, 1950, 1954, 1967; Rasch, 1968, 1980) and statistical interactions (Goodman, 1970).[2] These analytical innovations have shaped theory and hypothesis testing in behavioral and social sciences and,

[2] The history of social science inventions should become an important part of any sociology of knowledge as well as being integral to the study of social change. The ways that demand shapes intellectual agendas is not well understood. Consider the fact that Lazarsfeld undertook his work on latent structure analysis and Guttman on scale analysis in connection with research for the Research Branch of the Information and Education Division of the U.S. War Department in World War II.

as Holzinger noted in 1941, have had major applications in physics, medicine, and business forecasting (1941:5).

Sample Surveys Modern sample surveys rest on early inventions. The principles of random selection, objective probability, and stratified random sampling are well over a thousand years old (Duncan, 1984:iv). Survey modes of data collection also have been around for a considerable time. But the coalescence and systematization of these inventions into the modern stratified probability survey of a population are a product of modern behavioral and social science, coming mostly within the last 50 years.

As in the case of testing, there is a dearth of data to assess the effects of this invention, particularly its role in social change. Yet, we can plausibly argue that, except for institutional data collected as a by-product of organizational routines, the sample survey has become the major mode for linking action to intelligence in modern democratic societies. Even news organizations do not any longer claim to speak for the aggregate except in a metaphorical sense; but the opinion poll is accepted as doing so.

It is difficult to trace all of the ways that the sample survey has come to dominate organizational and individual decisions and operations. A few examples are offered simply to illustrate how pervasive it has become and how instrumental it is in changing behavior.

Perhaps nowhere has the invention of sample surveys altered the pattern of activity as much as in American electoral politics. Despite an abundance of skepticism about candidate and opinion polls, no candidate runs for major political office without a private polling operation. Media coverage of elections compares candidates in terms of their poll status; legislative and executive action is responsive to poll information; and political issue and candidate polls are a substantial American industry.

A second major area where surveys dominate is in providing intelligence for government decisionmaking. Much of the information for operating the government comes from sample surveys. The IRS, for example, has used sample surveys in its Audit Control Programs since 1948, and as an established part of its Taxpayer Compliance Measurement Program (TCMP) since 1962 (Long, 1980:55). These surveys of tax returns and filing compliance in the general population have become a principal means for the IRS to set its enforcement strategy. Major short-term policy indicators on unemployment and the cost of living are based wholly or in part upon sample surveys. The Survey Division of the Bureau of the Census has become one of its largest, quite apart from many other divisions within the bureau also operating sample surveys or collecting information through them. The Current Population Survey annually reaches about 1 in 1,000 households. No organization of any size remains unsurveyed by some government organization (though not always by sample surveys).

A third major area for sample surveys is marketing. Market research may be the dominant sector in sample surveying, surpassing the resources allocated to surveys by governments—though data for precise comparisons are lacking.

There are several kinds of market research. Sample surveys affect product development and sales strategies. They locate territories or populations for marketing a particular good or service. Surveys estimate the demand for new products or satisfaction with existing ones. The mass media, which rely on sample surveys for news, rely even more heavily on them for market information. No industry is more sensitive to the sample survey than television, in which ratings of network programs determine advertising revenues and the fate of writers, producers, and stars.

As a fourth major consequence, the sample survey has become the major means of developing social indicators in postindustrial society. Sample survey information is aggregated into indicators in two different, albeit related, ways. Surveys are used cross-sectionally—at a point in time—to evaluate relative performances or outputs, as in the Nielsen ratings of television programs, or to compare electoral candidate strengths. Social indicators are also used to forecast, monitor, control, or respond to the course of change over time. For example, the monthly Current Population Survey estimates unemployment, residential tenure, and vacancy rates; the semiannual National Crime Survey examines victimization rates; the Annual Housing Survey reports characteristics of housing units; and the National Health Survey examines illness, use of health care services, and health-related expenditures.

Sample surveys are also important in applied social science research, especially by nonacademic organizations. Not only has evaluation research become a substantial private industry, but major organizations such as the Armed Forces have developed a considerable in-house capability for sample surveys; it has been said that the most surveyed population in the world is the Armed Forces of the United States; certainly the American soldier in World War II served the most surveyed military in history (Stouffer et al., 1950).

Finally, the sample survey is one of the major methodological foundations of the modern behavioral and social sciences. Despite widespread use in government and by profit and nonprofit organizations, major innovations and inventions in sample surveying continue to stem mainly from the academic social science community. Exceptions occur, primarily in the development of efficient means of surveying, such as computer-assisted telephone interviewing (CATI); yet even when such innovations occur outside the we might discover that in postindustrial society behavioral and social science inventions are more consequential for social change than material inventions. Ogburn developed his theory of cultural evolution by focusing on the material inventions and advances in physical science and mathematics that contributed to the Industrial Revolution. That view scanted the great social inventions of earlier societies, such as bureaucratic administration and empires (Eisenstadt, 1963) and antedated most of modern behavioral and social science.[3] The role of economics in setting government policies and in the social control of economies has grown considerably since the work in *Recent Social Trends*. Although a president had sought the advice of academic social science in the "President's Research Committee on Social Trends," the committee seemed not to have imagined the significant role that behavioral and social science inventions would come to play in corporate organizational life and government in America.

Ogburn believed that the cultural base of social invention accumulated less rapidly in modern times than that of mechanical invention (Ogburn & Nimkoff, 1940:792).[4] This slower growth, in turn, slows the rate of new social invention. Yet there appears to be greater accumulation in the behavioral and social sciences than Ogburn expected. Rapid expansion of the knowledge base has been especially evident in cognitive psychology and linguistics.

A final word may be in order here on the reluctance to examine the impact of behavioral and social science inventions on society and especially on social change. Lags in adaptation due to such inventions may be intrin-

[3] Ogburn observes en passant: "The fact that technology is at present so powerful a cause of cultural lags, and consequent social disorganization, does not deny that other variables such as social inventions or population changes are creating lags also . . . the lag of social changes behind technological progress is simply a special case of the general phenomenon of unequal rates of change of the correlated parts of culture" (Ogburn & Nimkoff, 1940:893).

[4] The matter is empirical. It is not clear that the cultural base of social inventions cumulates any less rapidly in the modern world. Boulding (1978) argues that the homogenization of societies throughout the world may lead to less diversity in the cultural base and thus in the long run threaten the survival of culture.

tions. This strategy of theory construction and testing downplays the important ways that inventions occur and are diffused in society—most often other than by deliberate intervention—and promotes the false premise that invention and intervention are ordinarily successful in producing change, except where organizational resistance is powerful enough. The contrary seems to be the case. Most experiments and inventions fail, or succeed in producing entirely unintended effects. We may learn more about how to produce intended effects through social invention by looking to the unintended consequences of purposive social action (Merton, 1936).

Reduction of Cultural Lags

Although Ogburn subordinated the role of behavioral and social science inventions in causing cultural change,[7] he assigned to these sciences a special role in facilitating the *adaptation* of society to changing material culture (1934:166). Ogburn believed that the failure of institutions to adapt to advancing technology produced nearly all social maladjustment and disorgani-

[5] Ogburn (1957b:8–9) concluded that the study of social trends carries two major messages: "The first general message that knowledge of social trends brings to us is that there is much stability in society, even though there be a period of great and rapid social change. . . . The second lesson we learn from a knowledge of social trends is that there is a sort of inevitability about social trends. . . . It is difficult to buck a social trend. It may be slowed up a bit, but generally a social trend continues its course. . . . Success is more likely to come to those who work for and with a social trend than to those who work against it."

[6] Antipathy toward military institutions, for example, may account for a general neglect of how organizations may change quite rapidly and as a consequence of social inventions. In the history of race relations in the United States, for example, little attention is given to how the U.S. military organizations became egalitarian and at an accelerated rate compared with any other sector of American society (and that religious organizations are among the most recalcitrant to change and racially segregated at the local level).

[7] In Part VII, "Social Change," of *Sociology*, Ogburn recognized that assigning a priority to mechanical invention is partly a function of the precision with which an invention can be dated. He also recognized the problem of an infinite regress of causation that complicates assignment of priority in social change. He concluded with a mechanical analogy: "When all the interconnected parts of a culture are in motion, and each part exerts a force on some other part, the origin of the motion cannot be located" (Ogburn & Nimkoff, 1940:866–867).

mulation of lags was thus inevitable, it could still be reduced. For example, wars and revolutions reduce accumulated lags in a society (Ogburn, 1957a:172). Another less radical way to reduce lags is through the technology of the behavioral and social sciences (President's Research Committee on Social Trends, 1933:166). But just how to achieve this Ogburn failed to make clear.

The answer would have to lie in the production of knowledge-based innovation and invention designed to increase adaptation to cultural changes or to reduce the effects of their accumulation.

Below I will illustrate two different ways in which social science—both basic and applied—can function in restructuring societies in consequence of changes in culture.

Statistics and Quality Control The invention and diffusion of statistical quality control illustrates how social inventions can cope with the cultural dislocations caused by material and nonmaterial inventions. The coalescence of mechanical inventions into the modern mass production assembly-line factory produced the problem of assuring uniformity and high precision. Departures from strict production standards have consequences ranging from mechanical failure to increased transaction costs; these can be very significant in competitive markets or under other conditions where the tolerance for failure is small.

Statistical quality control is the statistical surveillance of repetitive processes. It is used primarily for two purposes: *process control* to evaluate future performance and *acceptance inspection* to evaluate past performance (Wallis & Roberts, 1956:495). In either type of control, samples are drawn to make decisions about a population. For process control, the population is an infinite number of expected results from repetitions of the same process; for acceptance inspection, it is the quality of a finite set of existing items.

The basic invention of statistical quality control was developed in the 1920s by an industrial statesman, Shewhart,[8] who invented the statistical quality control chart (1925, 1926a, 1926b, 1927, 1930, 1931). Its wide-

[8] Shewhart dates the invention of the statistical quality control chart as 1924 (1939:4).

spread dissemination came in the 1940s and resulted from the demands of the War Production Board, which deemed quality production of military goods essential to winning the Second World War, especially in light of the high quality of the German industrial complex (Wallis & Roberts, 1956:495, 512). Wald's method of sequential analysis (1945), although developed initially for use in scientific research, proved so useful for acceptance inspection that an estimated 6,000 U.S. plants used it within two years of its development in 1943 (Wallis & Roberts, 1956:518).

Other organizational innovation accompanied this rapid diffusion. Intensive training courses in quality control were developed at Stanford University and given in most major industrial centers during the war. Among the many consequences of diffusion was the founding of the American Society for Quality Control, made up largely of applied statisticians working in industrial applications.[9]

Ogburn concluded from his studies that the acceptance of inventions and their integration into cultures other than the one of origin depended upon the similarity of the cultures involved (Ogburn & Nimkoff, 1940:829). He was also disinclined to assign causal roles to individuals either in invention or diffusion (Ogburn, 1926). For Ogburn, the existence of independent invention demonstrated that the cultural base predominates over individual ability or uniqueness.

Ogburn's view may be correct in the long run, but in the short-run case of quality control, there were key individual disseminators. One of these was W. Edwards Deming, a government statistician originally in the Department of Agriculture and later at the Bureau of the Census and on independent government assignment. The introduction and rapid diffusion of statistical quality control in Japan seems largely due to the efforts of Deming. Since 1951, the Union of Japanese Scientists and Engineers has recognized his importance to Japanese industry by creating a major award, the Deming Prize, for contributions to statistical quality control in industry (American Statistical Association, 1983:1).[10] Some believe that the competitive margin of Japanese over U.S. products is attributable to a higher integration of statistical quality control in Japanese industry.

[9] Although statistical quality control was initially developed and applied in industry, the invention has wide applications since it is applicable to any kind of repetitive process, e.g., communicable diseases, medical experiments with human subjects, and accounting processes.

[10] There is no Deming Prize in the U.S., although he was honored in 1983 by the American Statistical Association for his contributions to "statistical quality control at home and abroad" with the Samuel S. Wilks Medal Award. Deming also has been decorated for his work in the name of the Emperor of Japan with the Second Order Medal of the Sacred Treasure. Nearing age 83, the peripatetic Deming was absent from the award ceremony, unable to fit it into his schedule without a few months' notice! (American Statistical Association, 1983:1).

Cohort Analysis A second example of how behavioral and social sciences permit adaptation to social change is the use of cohort analysis. A cohort is an aggregate of individuals of similar age who are exposed to or experience certain events during the same period of time. Cohort analysis is a quantitative description and analysis of occurrences from the time a cohort is exposed to these events (Ryder, 1968:546).

The continued entry of new cohorts provides a continuing opportunity to modify society. Cohorts consequently are central to the study of social change. But there also may be effects associated with age or aging per se, and changes brought about by external influences or events that affect all people alive at the time. These three sources of change in a population are referred to as cohort, age, and period effects.

A cohort analysis, as Ryder (1968:550) points out, differs from a longitudinal or panel analysis in that the latter examine changes in the individual members of a population or sample over time, while cohort analysis examines the changing characteristics of an aggregate through time: it is macro- rather than microlongitudinal.

The value of a cohort analysis to our understanding of social change can be illustrated by the studies of changing attitudes toward racial integration in the United States (Taylor et al., 1978:48). Opinion polls between the 1950s and 1980 showed considerable shift in white attitudes favoring racial integration. Underlying that shift, however, were different cohort trends. Although all age groups showed some shift with aging, this factor accounted for only about 10 percent of the total attitude change. Almost half of all change was due to the succession of cohorts in the population, with older, less favorable cohorts being replaced by new, more favorable ones. Almost half of the change in favorableness by 1980 is due simply to those younger cohorts comprising an ever greater portion of the population. By simple extrapolation we would forecast that within a matter of decades the vast majority of the population will favor racial integration. This type of cohort analysis shows that lag reductions often occur through the mechanism of population replacement.[11]

[11] But cohort analysis does not substitute for theoretical models of what causes particular changes. In the example, we still need to explain why the younger cohorts are most favorable. Is it due, for example, to indoctrination, to greater contact with unlike persons in environments such as schools, to involvement in social movements that support certain racial attitudes, or to some combination of these and other explanatory variables? While cohort analysis can aid us in understanding changes at the population level, it does not provide a substantive theoretical explanation of how such changes occur at the macrolevel of individual members of that population or at the microinstitutional and organizational level of changes. The failure to develop explanatory micro- and macromodels of social change severely limits our understanding of it. For a more extended

Cohort analysis has becom
social planning, especially w
casting. It has figured promin
the Social Security system,
prison and prisoner legislatio

Cohort analysis does not ne
the birthrate. One does not, f
estimate in 1980 the size of th
or assumptions about rates o
tioning are required. But giv
transitional probabilities, one
expected annual change in cr
age composition of a populati
example, demonstrated that t
pecially sensitive to changes
these characteristics of living
commitments in 1985, and th
apprised the Pennsylvania leg
sidering could have a consid
population in Pennsylvania p
crease in mandatory sentence
templated by the Pennsylvan
the prison population by 50 p
(1982) has shown that chang
appropriate procedures can b
size of prison populations. Bu
of time to complete, do littl

sions by estimating how cohort succession can be
market for old and new products.

CONSEQUENCES OF SOCIAL CHAN
MEASURES AND MEASUREME

Concepts and Measures as Products of

The social process itself is the source of most ba
cedures of social measurement (Duncan, 1984:ii). A
socially organized (Biderman & Reiss, 1967), and
into social life are subject to its substantive laws e
discover and test those laws (Reiss, 1980).

Duncan's recent work on social measurement (19
the fact that many of our basic concepts and proc
surement such as voting, counting people, money, sc
punishments, randomization, and sampling did not
of scientific knowledge but rather as the consequenc
solving. Not only do we depend upon social proces
our concepts and measures, but the development
social science depend in the long run, as Duncan
society wants or allows to be measured and is able
How it will be measured—or, in any event, the soci
concepts and measurement—is also socially determi
sicists and astronomers who struggle for appropriati
particle accelerators or space science vehicles face

Because behavioral and social
processes and require support i
statistical system to measure and
if it is firmly institutionalized.
know how difficult it is to find
over time, since repeating of q
interest and the salability of info
indicators often are eroded by so
Census of Manufactures was aba
miliar with attempts to institution
cators of attitudes and opinion kn
Research Center's General Socia
of support from the National Sc
series to span U.S. history are th
poses of apportioning democratic

Over-time measures not only de
and maintenance but may change
This is apparent in the developm
and legal regulation. The behav
processes such as legislation, ch
administration, the changing tec
practices. Major fluctuations in
Drug Administration from 1919
changing legislative mandates for
for inspectors, and changes in st
erman, 1980:229–241).[12]

may be no more litigious—perhaps even less so-
always have been litigated. Stated more technicall
and reliability are themselves grounded in postulat
measurement of change.

Social Change and the Organization of Ways o
surement of social life and changes in it depends u
ways of knowing. The concept of a "real" or "t
example, independent of organized efforts to detec
illusory (Biderman & Reiss, 1967). There are no rat
intelligence system to demand, collect, and proc
whether the people are scientific observers, polic
jurists. More generally, there are only socially orga
because all criteria (and measures) for knowing, d
social facts lie in social organization (Biderman &

Ogburn conceptualized the organization of ways
invention, and indeed our understanding is enhanc
as the product of invention. Many small inventions,
what we think of as the modern sample survey: n
as statistical probability, sampling, and analytical
zational procedures to train and supervise people i
in interviews and to link these people together in a

These observations have a number of implication
change. One is that we must understand how these
of knowing change over time as a consequence of n
technology and the effect that such changes hav

in socially organized and epistem
tion, for example, that a method
population can be exactly counte
by enumerating everyone who r
time.

The Paradox of Method All
erned by the substantive laws of
are the means for discovering an
paradox can be resolved only by
will advance the formulation an
ciprocally, improvements in sub
vances in method. The paradox
Kaplan's (1964:53–54) paradox
are needed to formulate a good t
at the proper concepts." This
approximation.

In exploring the study of socia
of the important implications
theories and the development
development of substantive know
another way, research on metho
development of all knowledge de
to the development of any scienc
germane to its methods and me

Some behavioral and social s

sources of variance (Webb et al., 1966:4). Here the po is inextricably woven with substantive theories abou lying the method. Even so, the assumption of multip multimethod approach (Campbell & Fiske, 1959; depends upon substantive theories and knowledge: it re that components can be weighted according to the variation and in combination by their independence f of bias. Indeed, one problem of a multimethod app why methods are not substantially of equal weight.[14

There seems to be no escape, then, from the par method. We must be prepared, consequently, to devo to understanding the substantive theories that underlie it can be argued that the most critical theory for the germane to its methods (Reiss, 1980).

We must therefore draw two more implications of t for the study of social change. First, concepts and vestigate social change are vulnerable to secular c methods of measuring and analyzing social change are changes. The problem for theorists and empirical in measure social change when both the measures and measured are changing. Or correlatively, how to m when that which is being measured is changing wh measures are too limited or rigid to detect it. The ta complex, but must be faced if we are to scientifica change.

Consequences of Institutionalizing Measure

The difficulties social scientists encounter in mea may preclude the kind of precision we commonly asso sciences. The past may never be kept in such a way the future, once we discover or invent new ways o Moreover, any evolutionary or dynamic theory of

Most indicators of change emerge
depends upon such processes. Cons
to at least two major types of change
of change affect the concept and its
below. On the other hand, the orga
will, from time to time—in response
its operational measures, and chang
substantively and operationally, or
of data collection and measuremen
occurring within the phenomena und
be accretions to that class of phenon

Examples of how concepts are re
observed phenomena are the repeate
the Census in the definition of a d
changes respond largely to the way
change from one decade to the next
household'' by the bureau is a resp
concept was biased toward older pe
tional procedures and conceptualiza
include ''secondary family workers'
Yet other seemingly simple census
include ''ethnic status'' (vulnerable t
tive language,'' ''country of origin,

Examples of increments to a clas
concepts. Congress, for example, n
on arson as a violent index crime in

This was conceptualized in terms of a *work force* com
occupied workers in the nation. Moreover, each wor
have a *usual* occupation, which was asked about as
information on labor resources (Jaffe & Stewart, 1
Labor resources were regarded as important to jobs,
status, not employment status, was usually measure
for example, reported statistics on the gainfully em
their usual occupation. There were no national statist
because this usual occupational status concept of th
include current employment status.

A conception of unemployment linked to the bu
course, existed for quite some time. But prior to 19
ployment was largely confined to labor unions. Unic
for systematic documentation. Unemployment durin
and depressions of the nineteenth and early twentiet
casional estimates of unemployment and, beginning in
concern with the effects of changing technology led
occupational composition, employment, and unemp
sectors undergoing rapid technological change. Ther
of seasonal unemployment in agriculture. But all of
based on a presumption that unemployment was a te

Consequently, there was surprisingly little statistic
employment in the chapter on labor in *Recent Soci*
Research Committee on Social Trends, 1933:xvi). T
system was unprepared to measure it. As Jaffe and
cluded, although *gainfully occupied* statistics may

people entering the work force
workers, and those who earned
erally omitted. Not regarded as
as unemployed when they sough
of the gainfully employed was s
injury, death, and emigration; it
of young people and immigratio
a usual occupation or a custom
consistent with a stable society b
ing urban economy with mobile
are unlikely to change in the sh
at long intervals such as the dec

The rapid growth of unempl
public interest in short-run unen
frequent estimates of unemplo
needed—jobs. This required a
dividuals to the labor market. Th
occupational status and experien
emerging concept was a labor fo
for work in the preceding month,
of being called back to work. Th
of individuals, and what emerge
force concepts, an increase in t
decrease in the unemployed. Und
are measured somewhat separa
employed and unemployed can

ties. The line between a person looking for work
vague, and the reasons for not looking for work we
the committee decided to exclude "discouraged"
had given up looking for work because they saw litt
job. In 1967, the Bureau of Labor Statistics implem
dation and excluded a substantial segment of the une
the concept of looking for work to those searching
preceding four weeks. At the same time it began
"discouraged" if they said they were not looking for
were there a job for them.

As recently as 1979, Finnegan advised the Nati
Employment and Unemployment Statistics (1979:21
practice of reporting discouraged worker statistics—n
rather than monthly—should be continued, this group
as unemployed. But the matter did not end there either
are disproportionally distributed among persons unc
60 years of age and among blacks and women. Sh
be regarded as in some way unemployed? Clearly,
long-run changes in the society are at the heart of
political debates.

Our example of unemployment highlights a centra
surement of social change—that the concepts, opera
measures used to chart this change are themselves
a result of social processes. What seems called for is
how one adjusts concepts and their measures over ti
sures, synthetic estimation, and multiple measures al

bers of a population keep their own statistics. Indeed, were an invento taken of statistical series, government series would probably contribute t minority.

Our concern in this section is with the consequences of the curre organization of statistical indicators for monitoring and investigating soc change. Perhaps the major consequence of the current organization of c statistical system is that we cannot readily compile them into meaning aggregates beyond the level at which they are collected. Judicial statist are gathered for every court in the land, but it is difficult to combine mo than a small part of them into state (much less national) statistical seri Conversely, most national statistics are collected in such a way that cannot make local estimates from them. This is especially true for ser collected by the survey method, but it is so for many other modes of da collection and compilation as well.

A second major consequence is that statistics gathered by private org nizations generally are inaccessible for aggregation or analysis unless porting is centrally coordinated and controlled. Private organizational da are seldom compiled unless there is legislation creating a voluntary mandatory system of reporting. As a consequence, we cannot measure ve much change using that vast resource.

A third consequence is that the major developmental and analytical sources are concentrated in federal statistical systems designed to me particular needs of federal legislative, executive, and judicial agencies. T resulting lack of attention to local variation may have the consequence th matters requiring collective attention as well as social changes come to defined for a national aggregate rather than in terms of their local variabili To the degree that statistical information is important in reaching decisio this imbalance may bias toward federal rather than state and local adaptatio Countries with central statistical systems such as France were historica organized to gather information for each department and unit thereof. might be useful to learn more about the role such regionalized statisti systems play in social change in contrast with the United States. I wou draw attention to the problem of developing concepts and measures of soc change that meet agreed-upon requirements of all levels of governme Volunteer sample surveys, rotating panels, and synthetic estimates are so of the ways of doing so. This is an area for social invention.

THE CONCEPTUALIZATION AND MEASUREMENT OF SOCIAL CHANGE

Most current theories and models of social change are deficient in planatory precision and predictive power. This section considers some the ways that contemporary models and concepts could be improved to better our understanding of social change.

Individualistic Biases in Studying Social Change

The dominant theories of social life in the United States postulate individuals as basic units and especially presume individual actors who make rational decisions. References to collective choice and organizational actors are often translucent, magnifying postulates about the behavior of individuals to collective actions. Durkheim's view of society as a reality *sui generis* is honored more in the breach, as interpretative commentary.

Ogburn was well aware of the domination of individualism in explaining social change, formulating the problem as the role of the "great man" versus "social forces" (Ogburn, 1926). His own well-known view was that great men are but *a medium* of social change (1926:231). Historical theories, such as those of Sorokin (1937–41, 1943, 1947), likewise assign a key role to social forces, which dominate the behavior of individual actors. But the contrary emphasis upon individual actors and individual welfare still dominates much contemporary theory and research, biasing the treatment of social change.

Earlier I used unemployment as an example of a major social indicator. Unemployment is a characteristic attached to persons. So is the concept of the discouraged worker, based as it is on motivation of individuals in a labor market. Even though it is apparent to labor economists that employment attaches to jobs, we lack indicators of job vacancies per se; there is no national indicator of jobs comparable to that on unemployed individuals.

In his ground-breaking studies of vacancy-chaining, Harrison White (1970) noted that social scientists had focused on social mobility as individual movement through jobs, neglecting the fact that a job, as a position in an organization, must open up or become vacant to constitute an opportunity for mobility. White studied movements of vacancies through bureaucratic hierarchies—how vacancies are filled and how the filling of positions relates to organizational structure and process. Positions and opportunities tend to be linked in chains, and these constitute the relative openness to upward mobility in a bureaucracy.[18] Although recent studies of social mobility have examined how organizational structures affect occupational mobility, defining these as opportunity structures (Rosenbaum, 1976, 1984), the study of mobility has not shifted to focus on changes in opportunity structures themselves.

To the extent that theories dictate what is problematic, they also dictate

[18] Parenthetically we note that the trickle-down market allocation theory of housing in economics could be tested in a vacancy-chaining model.

secure information about organizations. Indeed, mos methods for collecting information on organizations actually collect infor mation about individuals in organizations or rely on individual surrogate for the organization. Far too little attention is paid to measures other thar surveys (Sinaiko & Broedling, 1976).

Even when organizations are the object of study, the line of discussior is affected by the individualistic bias. For example, Kish (1965) noted tha around the time of the first Sputnik, about half of U.S. high schools offerec no physics; a quarter, no chemistry; and a quarter, no geometry. He then noted that this did not tell us how many high school students could take courses in chemistry, physics, or geometry, for the schools offering no such courses, though large in number, were small in size, accounting for only 2 percent of all high school students.

It is clear that average school characteristics would give a misleading description of conditions for the average student. But Kish failed to point out the relevance of the original organizational statistic for deploying collective resources. For some purposes the distribution of schools is critical, e.g., if a government sought to equalize educational opportunities for *all* students. To do that, the government would have to merge schools or divide resources among existing schools; and in either case, a large number of schools would be involved. Fifty percent of all high schools, for example, might need a new physics laboratory and an instructor trained to offer physics. As a consequence, the market for physics teachers might change drastically and organizational consequences on teacher recruitment and training would be considerable. One can readily imagine a whole train of organizational, structural, and individual consequences stemming from how one reads these statistics and decides to act on them.

We routinely conceptualize and measure the size and composition of populations of individuals but have only recently come to think seriously about doing so for populations of organizations.[19] Yet the size of the organizational population in the United States is greater than that of individ-

[19] Networks are even more complex. Consider that apart from unmarried siblings, the kinship network is never the same for any two individuals. For further discussion of research on organizations, see Hannan, in this volume.

uals. Consider that a household is a form of organization, that it is not coterminous with the family as a form of organization, and that these may be regarded as two distinct populations of organizations. Alexis de Tocqueville sensed the multiplicity of organizations when he characterized America as a society of joiners; yet he focused primarily on the individual characteristics of the joiners and less so on the fact that American society was creating an enormous number of organizations for individuals to join, many such organizations persisting well beyond the involvement or the life of those who founded them.

The suggestion here is that organizations may play as great, if not a greater, role in social change than individuals, and that the bias toward individualism fails to take into account how populations of organizations are both causes and effects of such change. This, indeed, is not the only consequence of the individual bias; other units are neglected as well, such as units of culture. It has perhaps seemed simpler to count individuals when faced with the difficulties of devising and counting organizational populations or cultural products.

Ogburn held that change in material culture—invention—is fundamental to social change. One test of his theory required counting the numbers of inventions so that one could calculate the rate of growth of the technological base. Although he used patents to count the growth of invention, he recognized the limitations of this indicator. It perhaps is unfortunate that he did so little to try to count *social* inventions.

Ogburn's neglect of how one conceptualized and counted social inventions should be seen in historical perspective. Societal intelligence on the growth of science and technology is little advanced over Ogburn's day. At the core of counting inventions are conceptual problems of what constitutes invention and of how one measures the growth of knowledge in the sciences.

Some 30 years ago Lazarsfeld and Barton (1951) wrote a piece on measurement in the social sciences for a volume edited by Lerner and Lasswell on the policy sciences. They drew attention to the fact that while the individual was a primary unit of observation and measurement, there are also units that are not based upon the primary characteristics of individual members precisely because no individual data correspond to them. A community can have a speed law; individuals cannot. A community can be characterized in terms of the proportion of its members who violate those laws. A primary characteristic of a unit, they noted, had to be distinguished from its analytic characteristics, which refer to component elements. Organizations may be seen in terms of their individual members or in terms of properties that cannot attach to individuals.

Although individual analytic characteristics may help explain social change, as, for example, the proportion of a society's manpower that is employed in science, individual scientists do not and cannot have a technological base. The neglect of the latter for the former data may account for our being in a disadvantageous position to theorize on and measure social change. Even in our use of cohorts—consider the examples used earlier in this paper—we usually look at cohorts of individuals, rarely at cohorts of organizations. Bankruptcy is generally expressed in annual rates rather than survival rates in a birth cohort of organizations.

The study of social change should focus, then, much more on the primary characteristics of organizations, which should be regarded more in terms of functional subunits than participating individuals. We can readily see that resources (such as laboratories) and relational properties (such as hierarchy) are primarily characteristics of organizations, not individuals. We must systematically collect better information on characteristics of organizations and units of material and nonmaterial culture, to use Ogburn's terms, if we are to understand cultural and social change.

Individual versus Collective Welfare There is a bias in welfare models of human behavior toward optimizing or maximizing individual welfare rather than the welfare of collectivities such as organizations.[20] Trade-offs commonly are seen in terms of individual rather than collective costs and benefits. The quality of life is measured in terms of individual rather than collective units: Is this community a good one for scientists rather than for science? Is the housing stock fit for individual habitation rather than what kind of collective life is possible, given the housing stock? How one asks the question can make a difference. We look at the crime rates of communities in terms of victimizations in a population of individuals, neglecting the high rate of victimization of organizations and collective property—parks, schools, playgrounds. We look at individual careers in crime rather than at community careers in crime (Reiss, 1982a, 1983); yet the latter may explain much of the former.

Concepts such as justice, social cohesion, and social integration are not reducible to the lives of a society's individual members, nor can they be measured simply by summing observations for individuals. Changes in particular social indicators can have collective and individual effects. A change in the divorce rate, for example, is both a change in the status of

[20] Measures of social welfare as conventionally defined in political economy should not be confused with measures of collective welfare. Measures of social welfare typically are based on the concept of a collective consensus based on individual preference scores. Although such preference measures may be technically infeasible, they presume that consensus measures on welfare preferences optimize or maximize collective welfare, which is an empirical matter.

individuals and a change in social relationships and organizational structure of society. Most divorces increase the number of single-person households and decrease the number of two-or-more-person households. Divorces alter the relationships of husband to wife, children to parents, insurers to insured, and the taxable income and legal status of the parties, to mention only a few consequences of changes in the divorce rate. Such changes may produce chain reactions. The divorce rate can have a substantial effect on the size and occupancy rate of the housing stock, which may affect the burglary rate (a crime against housing units rather than individuals per se).

Understanding social change would thus seem to require understanding of collective as well as individual welfare, and how changes in collective welfare are consequential for individual welfare. We may need to think more about the well-being of science in society, less about the consequences of science for the quality of individual life. Controversies over the risks of science must be viewed not only in terms of the risk to individuals, such as by gene splicing, but also of how the failure to do gene splicing research may affect the state of science in a competitive order of societies.

Lags in Measuring Social Change

Ogburn edited an annual series of the May issue of *The American Journal of Sociology* called *Social Changes in* [Year] from 1928 to 1935 and in May 1934, one entitled "Social Change and the New Deal." In the early volumes, Ogburn made clear that his purpose was not that of editing a conventional yearbook but rather "scientific analyses of social change . . ." (1929). The Great Depression, with its marked social changes, had consequences for the publication of his annual series. In his introduction to *Social Change in 1932* (1933:823–824) he observed:

> *The American Journal of Sociology* has itself been influenced by these economic changes, and a policy of retrenchment in the interests of economy has affected the size of this special issue. We have had to reduce the number of the articles, as it did not seem possible to reduce the length of the articles further and have them of any scientific merit. . . . In order to do this, some of the topics covered regularly in the annual "Social Change" issue have been omitted. . . . In some cases the omission of certain topics is not a particularly serious loss because extensive data are not always collected every year in sufficient volume to note significant changes, and a two year interval will show the changes most clearly. This is true, for instance, in the case of social legislation. Most of our state legislatures meet only once in two years.

The effect that changes in society can have upon its intelligence system is disclosed here rather dramatically.

A second issue is also evident here. With what frequency shall we collect measures of social change? Ogburn calls our attention to the fact that frequency of measurement is in part tied to the social processes themselves. Changes of some kinds—especially those he would have characterized as adaptive—are institutionalized, such as in the periodic meeting of legislatures. The response to change will determine in part the scale and frequency of measurement and thus the capacity of science to detect and measure such changes.

A third point was also mentioned briefly in Ogburn's introduction—the problem of lags in our intelligence on social change. He notes in particular the lag between an event, its measurement, and analytical understanding of it (1933:823–824):

> There are few aspects of our social life that have not been markedly affected by this most severe economic depression of modern times. The papers in this volume indicate many of these changes and their effects. The extremely dramatic events, which began in the latter part of February and reached a climax in the most extensive closing of the banks ever known, have particularly significant effects. These, however, are not recorded in this volume, which is restricted to 1932. Some time has to elapse after an event for the data to be collected and recorded so that it is possible to submit them to scientific analyses. News events are almost simultaneous, but there must be a lag before the scientific analyses can occur.

Here we see a major and continuing issue in conceptualizing, measuring, and monitoring social change—that of how our intelligence systems can be developed to collect information on events as they take place and how we can reduce the lag between collection of information and scientific analysis. We often fail to collect information rapidly that is essential to scientific analysis while, at the same time, far more information lies in our collection systems than we can process. How to resolve these problems is not altogether clear. A good theory helps, but data collection also depends upon social processes.

Forecasting and testing likewise depend upon these processes. Since Ogburn's day, great strides have been made in time-series analysis by developing forecasting models and identifying causal and "leading indicator" models. Despite the inevitability of some lags in analysis and model testing, more attention still needs to be given to short-run forecasting as well as long-run theories of social change. Our capacity to measure and monitor social change depends upon using and encouraging all processes that represent investments in knowing about and understanding social change. Economics in particular has both a macro theory of change and methods of short-run forecasting. The theory develops episodically and partly as a consequence of social change; such concepts as stagflation emerge to cope with the inadequacies of the theory and to fit it more closely to a world "out there." Economics may have progressed rapidly precisely because it forecasts. Forecasts that fail are crucial steps in learning about theories and discovering where their weaknesses lie. But that means we must build models of social life so that it can be forecast. Both demography and economics have done so and learned much by failed forecasts; other behavioral and social sciences might well take note. At this juncture of theory building on social change, social change itself is the way to theory building. Thus, theory testing and failed forecasts may be the best paths to scientific understanding of social change.

Need for a National Statistical System

Although one can easily demonstrate that our national indicators of economic and social change are more highly developed in some areas than was true on publication of *Recent Social Trends*, in many areas there has been little improvement. For example, we still have few national indicators of legal change, and we rely almost exclusively on ad hoc surveys to monitor changes in values and value practices such as religious belief and observance. This discontinuity and variability in indicator development, collection, and reporting amounts to a failure to develop the national statistical system Ogburn envisioned. Some of this is due to benign neglect by the social sciences since World War II of macro social change in advanced societies.[21] One of the requirements for developing and testing theories of social change is a set of concepts and their indicators measured over time, within the domain of a national statistical system.

I note three salient conclusions about requirements for a national statistical system:

First, we require research devoted to building explanatory models of social change in order to structure a national statistical system that can usefully measure and monitor this change.

Second, because of the limits of present models of social change and underinvestment in their development and testing, we generally lack data on potential explanatory variables for the trends that *are* monitored and

[21] World War II appears to have been a historical dividing point in the study of social change. In the postwar period, the fashion in studying social change shifted to the "third world," the "developing nations," and "economic and cultural development." Modeling efforts shifted to how one might simulate the growth of economies and, increasingly for the noneconomic sciences, to the effects of rapid social change on traditional cultures. Although this latter interest fell within the domain of Ogburn's lag formulation, its shortcomings (see Smelser, in this volume) failed to generate interest in revising the model.

Third, we lack an adequate system of indicators of science and technology. The annual report of the National Science Board, *Science Indicators* (1983), has almost no major indicators to monitor substantive changes in science and technology, much less a set of explanatory variables related to such changes. Most glaring is the absence of indicators for the behavioral and social sciences and technologies based on them. The failure to collect information on the content of science and technology, especially on inventions, has two important consequences. One is that we do not measure changes in the rate of behavioral and social science inventions and technology and their contribution to contemporary society. The other, and more important, is that we are unable to measure relative contributions to social change, especially the contributions of behavioral and social science and technology compared to that of physical science and technology.

A SUMMING UP

The decades since the publication of *Recent Social Trends* have been a period largely of benign neglect by the behavioral and social sciences in modeling and measuring social change, economics being the major exception. This neglect may owe in part to the reticence of theorists, save for economists, to address matters of social change. But scientific knowledge shapes and is shaped by such change; it becomes practically meaningful in the context of what kind of change is, or seems, possible; and it is tested against these consequences. It may be well to remember that Ogburn, as Duncan (1964:vii) calls to our attention, ''saw science primarily as an accumulation of knowledge, but an accumulation whose structure is subject to continual change as new relationships among its parts are perceived or as discoveries shed new light on supposed relationships.''

Perhaps the period of benign neglect is drawing to a close, in which

[22] In acknowledging this, attention is also drawn to the dissatisfaction with the precision of measures of variables estimated in structural equation models of the economy or of leading indicators (see Klein, in this volume).

event it is essential to attend to the kinds of problems touched upon in this essay. We must understand better the role of behavioral and social science knowledge and inventions in social change. We must examine the effects of social change on theory, concepts, and measures, including their capacity to record and render social change intelligible. Time has favored Ogburn's conviction that statistical intelligence systems have a critical role to play in the processes of science and in society as a whole.

* * *

I wish to thank Otis Dudley Duncan and Barbara Laslett for their helpful substantive comments.

REFERENCES

Adams, R., Smelser, N., and Treiman, D., eds.
1982 *Behavioral and Social Science Research: A National Resource*. Washington, D.C.: National Academy Press.

American Statistical Association
1983 *AMSTAT News*, Number 100.

Bancroft, G.
1979 Some problems of concept and measurement. In National Commission on Employment and Unemployment Statistics, *Counting the Labor Force: Readings in Labor Force Statistics, Appendix Vol. III*. Washington, D.C.: U.S. Government Printing Office.

Biderman, A. D., and Reiss, A. J., Jr.
1967 On exploring the "dark figure" of crime. *The Annals* 374:1–15.

Binet, A., and Simon, T.
1905 Méthodes nouvelles pour le diagnostic du niveau intellectuel des anormaux. *L'Année Psychologique* XI:191–244.

Blumstein, A.
1983 Prisons: population, capacity, and alternatives. Pp. 229–250 in Wilson, J. Q., ed., *Crime and Public Policy*. San Francisco: ICS Press.

Blumstein, A., Cohen, J., and Miller, H. D.
1980 Demographical disaggregated projections of prison populations. *Journal of Criminal Justice* 8:1–26.

Boulding, K. E.
1978 *Ecodynamics: A New Theory of Societal Evolution*. Beverly Hills, Calif.: Sage.

Burt, C.
1915 General and specific factors underlying the primary emotions. *Report of the British Association for the Advancement of Science* 85:694–696.
1939 The factorial analysis of emotional traits. *Character and Personality* 7:238–254, 285–299.

Campbell, D. T., and Fiske, D. W.
1959 Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin* 56:81–105.

Cattell, J. McK.
1890 Mental tests and measurements. *Mind* XV:373–380.

Cook, E.
1914 *The Life of Florence Nightingale*. London.

Duncan, O. D.

1964 Introduction. *William F. Ogburn On Culture and Social Change: Selected Papers*. Chicago: University of Chicago Press.

1984 *Notes on Social Measurement: Historical and Critical*. New York: Russell Sage Foundation–Basic Books.

Eisenstadt, S. N.

1963 *The Political Systems of Empires: The Rise and Fall of Historical Bureaucratic Societies*. Glencoe: The Free Press.

El-Korazty, M. N., Imrey, P. B., Koch, G., and Wells, H. B.

1977 Estimating the total number of events with data from multiple-record systems: a review of methodological strategies. *International Statistical Review* 45:129–157.

Ericksen, E. P., and Kadane, J. B.

1983 Estimating the Population in a Census Year: 1980 and Beyond. Carnegie-Mellon University Technical Report #260, Pittsburgh, Pennsylvania.

Fienberg, S. E.

1972 The multiple recapture census for closed populations and incomplete 2k contingency tables. *Biometrika* 59:591–603.

Funkhouser, H. G.

1937 Graphical representation of statistical data. *Osiris* 3:319–344.

Galton, F.

1869 *Hereditary Genius: An Inquiry into Its Laws and Consequences*. New York: Horizon Press (1952).

Goodman, L. A.

1970 The multivariate analysis of qualitative data: interactions among multiple classification. *Journal of the American Statistical Association* 65:226–256.

Greenwood, P.

1982 *Selective Incapacitation*. Santa Monica, Calif.: The Rand Corporation.

Guttman, L.

1950 The problem of attitude and opinion measurement; Basis for scalogram analysis; and The principle components of scale analysis. Pp. 46–90, 271–272, and 312–361 in Stouffer, S. A., et al., eds., *Measurement and Prediction*, Vol. 4 of *Studies in Social Psychology in World War II*. Princeton, N.J.: Princeton University Press.

Holzinger, K. J.

1930 *Statistical Resume of the Spearman Two-Factor Theory*. Chicago: University of Chicago Press.

1931 On factor theory. *Conference on Individual Differences in Special and General Abilities*. Washington, D.C.: National Research Council.

1941 *Factor Analysis*. Chicago: University of Chicago Press.

Hotelling, H.

1933 Analysis of a complex of statistical variables into principle components. *Journal of Educational Psychology* 24:417–441, 498–520.

Jaffe, A. J., and Stewart, C. D.

1951 *Manpower Resources and Utilization: Principles of Working Force Analysis*. New York: John Wiley.

Kaplan, A.

1964 *The Conduct of Inquiry*. San Francisco, Calif.: Chandler.

Kelley, T. L.

1928 *Crossroads in the Mind of Man: A Study of Differential Mental Abilities*. Stanford, Calif.: Stanford University Press.

1935 *Essential Traits of Mental Life*. *Harvard Studies in Education* 26. Cambridge, Mass.: Harvard University Press.

Kish, L.

1965 Sampling organizations and groups of unequal sizes. *American Sociological Review* 30:564–572.

Klineberg, O.

1933 Mental tests. *Encyclopaedia of the Social Sciences* X:323–329. New York: Macmillan.

Lazarsfeld, P. F.

1950 The logical and mathematical foundation of latent structure analysis; The interpretation and computation of some latent structures. Pp. 362–472 in Stouffer, S. A., et al., eds., *Measurement and Prediction*, Vol. 4 of *The American Soldier: Studies in Social Psychology in World War II*. Princeton, N.J.: Princeton University Press.

1954 A conceptual approach to latent structure analysis. Pp. 349–387 in Lazarsfeld, P. F., ed., *Mathematical Thinking in the Social Sciences*. Glencoe, Ill.: The Free Press.

1967 *The Uses of Sociology*. New York: Basic Books.

Lazarsfeld, P. F., and Barton, A. H.

1951 Qualitative measurement in the social sciences: classification, typologies, and indices. In Lerner, D., and Laswell, H., eds., *The Policy Sciences*. Stanford, Calif.: Stanford University Press.

Levine, D. B., Hill, K., and Warren, R., eds.

1985 *Immigration Statistics, A Story of Neglect*. Panel on Immigration Statistics, Committee on National Statistics, National Research Council. Washington, D.C.: National Academy Press.

Long, S. B.

1980 *The Internal Revenue Service: Measuring Tax Offenses and Enforcement Response*. Washington, D.C.: National Institute of Justice (September).

Merton, R.K.

1936 The unanticipated consequences of purposive social action. *American Sociological Review* 1:894–904.

National Commission on Employment and Unemployment Statistics

1979 *Counting the Labor Force*. Washington, D.C.: U.S. Government Printing Office.

National Science Board

1983 *Science Indicators 1982: An Analysis of the State of U.S. Science, Engineering, and Technology*. Washington, D.C.: National Science Board and National Science Foundation.

Nightingale, F.

1858 *Notes of Matters Affecting the Health, Efficiency, and Hospital Administration of the British Army*. London: Harrison and Sons.

Ogburn, W. F.

1912 *Progress and Uniformity in Child-Labor Legislation: A Study in Statistical Measurement*. Studies in History, Economics, and Public Law 48(2). New York: Columbia University Press.

1914 Methods of direct legislation in Oregon. *Quarterly Publication of the American Statistical Association* XIV.

1915 Initiative and referendum tested in hard times. *Survey* 23:693–694.

1919 Measurement of the cost of living and wages. *Annals of the American Academy of Political and Social Science* VIII:235–242.

1922a Bias, psychoanalysis, and the subjective in relation to the social sciences. *Publications of the American Sociological Society* XVII:62–74.

1922b *Social Change: With Respect to Culture and Original Nature.* New York:
Press.

1926 The great man versus social forces. *Social Forces* V:225–231.

1928– *Social Changes in . . . (1928–1935).* Annual series of the May issue of *The*
1935 *ican Journal of Sociology.* Reprinted annually by University of Chicago.

1929 *Social Changes in 1928.* Chicago: University of Chicago Press.

1933 Social changes in 1932: introduction. *The American Journal of Sociology* 3
824.

1934 Trends in Social Science. *Science* LXXIX:257–262.

1942 Man and his machines: teaching American youth how invention changes the
world. *Problems in Modern Life,* no. 3. Washington, D.C.: National Coun
Social Studies, National Association of Secondary School Principals, Departm
the National Education Association.

1950 *Social Change.* New York: Viking Press.

1957a Cultural lag as theory. *Sociology and Social Research* XLI:167–173.

1957b Social trends. *Sociology and Social Research* XLII:3–9.

Ogburn, W. F., and Goltra, I.

1919 How women vote. *Political Science Quarterly* 34:175–183.

Ogburn, W. F., and Nimkoff, M. F.

1940 *Sociology.* New York: Houghton Mifflin.

President's Research Committee on Social Trends

1933 *Recent Social Trends in the United States.* New York: McGraw-Hill.

Prosser, W. L.

1964 *Handbook of the Law of Torts.* 3rd ed. St. Paul: West.

Rasch, G.

1968 An individualistic approach to item analysis. In Lazarsfeld, P. F., and Henry,
eds., *Readings in Mathematical Social Science.* Cambridge, Mass.: MIT Pres

1980 *Probabilistic Models for Some Intelligence and Attainment Tests* (with a new fo
and afterword by Benjamin D. Wright); expanded edition, University of Chicago

Reiss, A. J., Jr.

1968 Stuff and nonsense about social surveys and observation. Pp. 351–367 in Beck
Geer, B., Reisman, D., and Weiss, R., eds., *Institutions and the Person.* C
Aldine.

1971 Systematic observation of natural social phenomena. Pp. 3–33 in Costner, H
Sociological Methodology.

1976 Systematic observation surveys of natural social phenomena. Pp. 123–141 in S
W., and Broedling, L. A., eds., *Perspectives on Attitude Assessment: Surve
Their Alternatives.* Champaign, Ill.: Pendleton Publications.

1980 Exploring the central paradox of method in social science inquiry. Yale Uni
(Copyright 1980), New Haven, Connecticut.

1982a How serious is serious crime? *Vanderbilt Law Review* 35:541–585.

1982b Statistical measurement of social change. *The 5-Year Outlook on Science an
nology 1981* 2:649–667. Washington, D.C.: The National Science Foundation,
Materials.

1983 Crime control and the quality of life. *American Behavioral Scientist* 27:43–5

Reiss, A. J., Jr., and Biderman, A. D.

1980 *Data Sources on White-Collar Law-Breaking.* Washington, D.C.: National I
of Justice (September).

Rosenbaum, James

1976 *Making Inequality: The Hidden Curriculum of High School Tracking.* New
Wiley.

1984 *Career Mobility in a Corporate Hierarchy*. Orlando: Academic Press.

Ryder, N. B.

1968 Cohort analysis. *International Encyclopedia of the Social Sciences* 2:546–550. New York: Macmillan & The Free Press.

Shewhart, W. A.

1925 The application of statistics as an aid in maintaining quality of a manufactured product. *Journal of the American Statistical Association* XX (December), New series no. 152: 546–548.

1926a Correction of data for errors of measurement. *The Bell System Technical Journal* 5:11–26.

1926b Quality control charts. *The Bell System Technical Journal* 5:593–603.

1927 Quality control. *The Bell System Technical Journal* 6:722–735.

1930 Economic quality control of manufactured product. *The Bell System Technical Journal* 10:364–389.

1931 *Economic Control of Quality of Manufactured Product*. New York: D. Van Nostrand.

1939 *Statistical Method from the Viewpoint of Quality Control* (with the editorial assistance of W. Edwards Deming). Washington, D.C.: The Graduate School, U.S. Department of Agriculture.

Shiskin, J.

1976 Employment and unemployment: the doughnut or the hole? *Monthly Labor Review* 99:3–10.

Sinaiko, A. W., and Broedling, L. A.

1976 *Perspectives on Attitude Assessment: Surveys and Their Alternatives*. Champaign, Ill.: Pendleton Publications.

Sorokin, P. A.

1937–41 *Social and Cultural Dynamics*. 4 vols. New York: American Book Co.

1943 *Sociocultural Causality, Space, Time*. Durham: Duke University Press.

1947 *Society, Culture and Personality: Their Structure and Dynamics*. New York: Harper & Bros.

Spearman, C.

1904 General intelligence, objectivity determined and measured. *American Journal of Psychology* XV:201–293.

1927 *The Abilities of Man*. New York: Macmillan.

Spearman, C., and Holzinger, K. J.

1925 Note on the sampling error of tetrad differences. *British Journal of Psychology* XVI: 86–89.

Stamp, J.

1937 *The Science of Social Adjustment*. London: The Macmillan Company.

Stouffer, S. A., et al.

1950 *The American Soldier, Studies in Social Psychology in World War II*, Vol. I, *Adjustment During Army Life*, and Vol. II, *Combat and Its Aftermath*. Princeton, N.J.: Princeton University Press.

Taylor, D. G., Sheatsley, P. B., and Greeley, A. M.

1978 Attitudes toward racial integration. *Scientific American* 238:42–49.

Thompson, J. D.

1967 *Organizations in Action*. New York: McGraw-Hill.

Thurstone, L. L.

1931 Multiple-factor analysis. *Psychological Review* 38:406–417.

1935 *The Vectors of Mind*. Chicago: University of Chicago Press.

Tuma, N. B., and Hannan, M. T.

1979a Approaches to the censoring problem in analysis of event histories. Pp. 209–240 in

Schuessler, K., ed., *Sociological Methodology*. San Francisco: Jossey-Bass.

1979b Dynamic analysis of event histories. *American Journal of Sociology* 84(January):820–854.

Wald, A.

1945 *Sequential Analysis of Statistical Data: Applications*. Statistical Research Group, Columbia University. New York: Columbia University Press.

Wallis, W. A., and Roberts, H. V.

1956 *Statistics: A New Approach*. New York: The Free Press.

Webb, E. J., Campbell, D. T., Schwartz, R. D., and Sechrest, L.

1966 *Unobtrusive Measures: Nonreactive Research in the Social Sciences*. Chicago: Rand McNally.

White, H.

1970 *Chains of Opportunity*. Cambridge, Mass.: Harvard University Press.

Wigdor, A. K., and Garner, W. R., eds.

1982 *Ability Testing: Uses, Consequences, and Controversies*. 2 vols. Committee on Ability Testing, National Research Council. Washington, D.C.: National Academy Press.

Uncertainty, Diversity, and Organizational Change

MICHAEL T. HANNAN

> There is in our social organizations an institutional inertia. . . . Unless there is a speeding up of social invention or a slowing down of mechanical invention, grave maladjustments are certain to result. (President's Research Committee on Social Trends, 1933:xxvii.)

How difficult is it to reshape complex organizations when conditions change? Ogburn's (1933) work on technical innovation and society built on the premise, illustrated in the quote above, that organizations and social institutions strongly resist change. He argued that the combination of rapid technical innovation and organizational inertia disturbs equilibria. Long periods of disequilibrium caused by lags in adjustment of social structures to changing material conditions can have high social costs, as Ogburn and his collaborators insisted 50 years ago.

Despite the seeming ubiquity of organizational inertia in everyday life, the social science literature has sometimes painted a very different picture. Both organizational theorists and specialists in management have often described a world in which organizational adjustment to changing external conditions is almost friction-free. March's (1981:563) review of the literature on organizational change notes this dominant theme:

> Organizations are continually changing, routinely, easily, and responsively, but change within organizations cannot be arbitrarily controlled. . . . What most reports on implementation indicate . . . is not that organizations are rigid and inflexible, but that they are impressively imaginative.

Which is it? Are organizations subject to strong inertial pressures as Ogburn has it? Or do they change easily and routinely as March claims?

The work reported here was supported by National Science Foundation Grants SES-8109382 and ISI-828013.

This disagreement raises fundamental issues about the relation of organizations and society, issues that have important theoretical and practical implications.

If change in organizational strategies and structures is rapid and smooth, it is reasonable to respond to changing conditions by continually redesigning existing organizations. But if organizations typically respond slowly or not at all to changing opportunities and threats in their environments, it may make more sense to continually replenish the stock of organizations. These alternative strategies imply vastly different social policies.

The disagreement between Ogburn and March reflects more than a generational shift in organizations theory and research. Contemporary opinion among organizational researchers splits sharply on these issues. Questions of organizational inertia are fundamental to understanding organizational structure and change; thus the opposing opinions voiced by Ogburn and March, which continue to divide current researchers, provide a useful framework for considering past and present theory and research on organizational change.

Inertia is only one of several noteworthy factors affecting the adaptability of organizations to environmental uncertainty. At least as important is the diversity of organizational forms in society—the stock of organized solutions to problems of producing collective action in variable settings. Trends that eliminate organizational diversity lower the capacity of social systems to deal with uncertain environmental change.

Questions about diversity and inertia are especially important when change in technical, social, and political environments is uncertain. If environments are highly stable (and thus certain), there is really no continuing problem of organizational adaptation. It will become clear eventually which forms of organizations are well suited to the stable, prevailing conditions (either by differential selection or by learning and imitation). Likewise, if environments change in predictable ways (for example, seasonal changes in demand for energy, Christmas trees, and other commodities), even highly inflexible organizations can schedule adjustments far enough in advance to match strategy and structures to these changing states.

Issues of organizational inertia and organizational diversity are important to understanding modern social change. This essay describes the development of theory and research on organizational processes as it bears on these questions. It also suggests new lines of inquiry that might better clarify the relations between organizational change and large-scale social change. In particular, it discusses recent theory and research that consider organizational diversity and change from ecological and evolutionary perspectives.

CENTRALITY OF ORGANIZATIONAL PROCESSES IN LARGE-SCALE SOCIAL CHANGE

Most theories in the social sciences emphasize the actions of autonomous individuals, interest groups, social classes, and institutions rather than those of concrete organizations. But almost all modern collective action takes place in organizational contexts; organizations are the main actors in modern society (Coleman, 1982). When interest groups and social classes take collective action, they do so using specific organizational tools such as labor unions, political parties, or terrorist groups. Recent research shows that even relatively amorphous social protest movements have a higher likelihood of success if they can use existing organizations (Tilly, 1978). The state, which has become the focus of so much social action, is itself an organization (or perhaps a hierarchy of organizations). Struggles for power and control in modern societies typically involve struggles between competing organizations for privileged positions in the state structure as well as struggles between the state and other kinds of organizations.

Organizations are also important in modern societies because of the role they play in creating, promulgating, and enforcing social norms. The codification of norms as explicit, legitimated organizational rules gives such rules great force. Organizations typically develop formalized roles and procedures for enforcing these rules. Employment contracts, for example, have more continuous and binding effects when labor unions monitor compliance and take action to enforce them.

Because organizations are key actors in modern society, the speed and direction of large-scale social change are constrained by organizational dynamics. In particular, the responsiveness of society to changing conditions depends on the inertia of its constituent organizations and on the diversity of its stock of organizations.

The problem of matching outputs of schools to the needs of a changing economy illustrates the problem. It has long been evident that American school systems were failing to teach enough mathematics and science to all but the richest and most able students. Over the past several years a series of national commissions has identified this situation as a "national problem" and urged immediate and far-ranging reforms of U.S. public education. Some commissions urge more attention to teaching (and requiring) more mathematics, science, and computing; others emphasize attention to writing. All agree that the quality of teachers needs to be upgraded and that more time must be allocated to teaching. A broad consensus seems to have emerged on the definition of the problem; federal and state officials, school district officials, legislators, school employee unions, and parents' groups all urge reform.

How quickly can the national system of public education be reformed? Despite the fact that many states have imposed new rules and constraints on school systems, there are a number of reasons for suspecting that change in the actual organization of schooling will be halting at best. The demographic and institutional constraints on change in this system are very powerful. Consider the problem of upgrading the technical knowledge of teaching staffs. In a period of declining enrollments, school staffs have been shrinking (although there has recently been an upsurge in demand for science teachers). Given the "last-in, first-out" policies favored by bureaucracies and demanded by teachers' unions, change in the composition of teaching staffs will be glacially slow without some radical alteration of employment policies. Any such radical change is sure to encounter stiff resistance from unions, as well as legal challenges. A radical change in policy may also mobilize previously quiescent groups.

The complexity of the organizational networks involved compounds the adjustment problem. There is no unitary chain of command; rather there are multiple, partially overlapping jurisdictions of local, state, and federal agencies, with no central planning mechanism. Change in any one sector is hampered by overlaps with others. For example, the seemingly simple problem of changing textbooks in public school systems is made very complicated by the organizational arrangements. Many different organizations and individuals must be consulted; any one of them can forestall the change.

Implementing even a broad and powerful mandate to change the educational system means changing many organizations and their interlocking connections. The whole system responds only as fast as the slowest component organizations.

Similar issues arise in industry, although the processes are different. In recent years, a number of highly concentrated American industries such as steel, automobiles, and agricultural and construction machinery, have stumbled before more efficient foreign competitors. The giant American firms in these industries adapted their strategies and structures to earlier technical and social conditions, and they have been ponderous at best in responding to new challenges. Firms in these industries have relied on political muscle to obtain favorable government intervention to limit competition, as we have seen in the auto and steel industries. Success in this tactic serves mainly to further delay radical change in industrial strategies and structures.

Global national policies like "reindustrialization" imply massive change in the structures of thousands of organizations. Whether such policies can proceed quickly enough to meet international competition and rapidly changing technologies depends largely on the responsiveness of existing firms in the economy and on the rate at which new firms can be created and brought p to speed. Analysis of such policies requires knowledge of the dynamics f organizational populations.

The discussion to this point has considered organizations as passive brakes n social changes initiated elsewhere in the society or in the environment. ut the image of organizations as passive is seriously misleading. Of course, rganizations *are* constructed as tools for specific kinds of collective action. or example, agents invest resources in hospitals or armies in the hope of chieving specific kinds of performances. But one of the main contributions f organization theory and research has been to show that organizations are ır more than simple tools.

Organizations consume great quantities of resources in merely maintainıg their structures. Because great quantities of resources are used for rganization building and for bureaucratic or administrative overhead rather ıan for production or for collective action, organizational politics often evolve around issues of resource allocation (Cyert and March, 1961; Pfefer, 1981). Organizational politics makes problematic the relation between echnical needs for production and actual distribution of resources. Subunits rive to protect and expand budgets and staff sizes. The resulting competition for fixed resources is especially severe in times of contraction or ecline (Freeman and Hannan, 1975; Hannan and Freeman, 1978). Because llocations within organizations are subject to intense political contest, rganizational action depends on the dynamics of political coalitions. Oranizational politics often makes collective action deviate from ostensible oals, from the demands of relevant environments, and from the intentions f organizational leaders.

For these various reasons attempts at understanding patterns of largecale change in modern societies (or relations between public policy and ctual implementation) require detailed attention to organizational processes nd dynamics.

Rational and Natural System Perspectives

Systematic organizational theory began when bureaucratic forms gained scendency as ways of organizing the activities of the state and of large ıdustrial concerns. German sociologist Max Weber, the founder of socioogical organizational theory, emphasized the importance of the spread of ureaucracy to the spread of norms of rationality. Bureaucracy, which is uilt on formalized rules, explicit spheres of competence, and full-time rofessional staff, permits rapid, efficient, and calculable response to adıinistrative directives. In Weber's (1978:973) view,

> he decisive reason for the advance of bureaucratic organization has always been purely echnical superiority over any other form of organization. The fully developed bureau-

> cratic mechanism compares with other organizations exactly as does the machine with the non-mechanical modes of production.

Because it is precise and efficient and because it can (in principle) serve the interests of any who come to control it, bureaucracy is practically indestructible in Weber's view.

Weber's insistence on the machinelike character of modern bureaucratic forms was echoed in this country by Frederick Taylor, the founder of the school of organizational design called Scientific Management (see Perrow, 1979, for a detailed examination of this school). Taylor described smoothly functioning organizations in which all tasks were broken down into minute components according to the logic of "time and motion studies." Research in this tradition sought to learn optimal designs for such organizational machines. Much work in this tradition, for example, tried to discover the optimal "span of control" for industrial organizations, the ratio of supervisors to workers that maximizes efficiency.

Much subsequent work in sociology, industrial psychology, industrial engineering, and economics was shaped by the broad assumption that organizations are efficient, impersonal tools for production, administration, and othe forms of collective action. Scott (1981) provides a careful summary and critique of work in this "rational-systems" perspective. This approach has produce much useful knowledge, especially about the conditions under which forma organizations have efficiency advantages in coordinating complex work. Man empirical findings of this tradition have become the conventional wisdom o management and public administration theory.

This perspective also continues to guide much current research. Fo example, an important development in economic theory of organization argues that organizations are often able to minimize the costs of completin economic transactions when markets fail due to imperfect information cognitive limitations on the ability to process information, and opportunisn (Arrow, 1974; Williamson, 1975).

Although the rational-systems perspective continues to shape research o organizations, most sociological research has long made an opposing argument. As early as 1915 German sociologist Robert Michels, who agree with Weber that bureaucratic forms were indispensable for efficient collective action, argued that bureaucracies seldom pursue their ostensibl goals. He claimed that organizations are subject to an "iron law of oligarchy." An organization requires expert leadership even when it is designed for democratic and collective ends, as in the case of labor unior and political parties. As leaders learn skills of managing and become differentiated in prestige and lifestyle from the mass membership, they develo interests in preserving the organization (and their privileged position) ny cost. They also develop the capacity to control organizational decisions. hus, Michels argued, leaders typically can and do subvert the goals of the rganization to minimize the risk that the organization will be destroyed. ccording to Michels (1962:364–365),

> . . the principal cause of oligarchy in democratic parties is to be found in the technical ndispensability of leadership. . . . Reduced to its most concise expression, the funamental sociological law of political parties (the term "political" here being used in s most comprehensive significance) [is] "It is organization which gives birth to the omination of the elected over the electors, of the mandatories over the mandators, of e delegates over the delegators. Who says organizations, says oligarchy."

Michels described a process by which an organizational tool takes on a life f its own. One result is that organizational action becomes highly unprelictable from knowledge of public goals and interests of its members. This nsight has been amplified by numerous studies in the so-called "natural-ystems" perspective (Scott, 1981), which stresses the continuities between ormal organizations and communities (Parsons, 1960; Selznick, 1948). Like ommunities, organizations have rich and complex political systems, and rganizational action is often the outcome of political contests among factions. ubunits of organizations seek to defend self-interests and resist reallocations f resources when conditions change. Moreover, members often develop shared orms in opposition to management. For these reasons organizations are at est "recalcitrant tools," as Selznick (1948) put it.

Much early work in the natural-systems perspective involved close exmination of the actual process of work in organizations, as in the famous tudies from the Hawthorne experiment at the Western Electric works Roethlisberger and Dickson, 1939). Also important were the case studies y students of Robert Merton at Columbia such as Peter Blau's (1955) nalysis of patterns of exchange in a social work agency and Philip Selzick's (1949) study of the relations between the Tennessee Valley Authority nd its local community. Recently, organizational sociologists have exended this perspective by conducting comparative quantitative analysis of rganizational politics. One particularly useful line of work, which follows he lead of the so-called Carnegie School (especially Cyert and March, 963), explores how control over essential resources converts to power vithin organizations and how power balances shape strategy and structure see especially Pfeffer, 1981; Pfeffer and Salancik, 1978).

As in the Weberian tradition, the natural-systems perspective has proided detailed empirical information about the limitations of organizational olutions to problems of collective action. It has identified the processes hat distinguish organizations from machines and shifted attention away rom idealized images of organizations and toward recurrent patterns of real organizational action. Numerous findings from this research tradition have also become enshrined in the conventional wisdom of management.

Theory and research on organizations during the past 20 years have sought increasingly to synthesize elements of the two starkly opposed perspectives. This work retains the premise of rational-systems theory that organizations are created as tools for collective action and that, in the long run at least, performance matters. That is, this synthetic perspective takes issue with the implicit claim of the natural-systems perspective that organizations are somehow shielded from negative consequences of inferior performance. It also rejects the naive claims of the rational-system perspective that organizations are simple, calculable machines. Instead it treats organizations as open systems that depend on a continuing flow of resources from environments. The necessity to maintain such a flow exerts at least some discipline on organizations. However, the fact that one essential resource—membership—comes with special interests and with attachments to other parts of the social world creates conditions of recalcitrance and inertia. According to various open-system perspectives, organizations are subject both to environmental constraint and to strong inertia. The main theoretical problems concern the relation of these two kinds of constraints.

These issues are most interesting theoretically and most relevant to practical problems when they are considered in the context of organizational *change*. Despite the fact that inertial tendencies seem to be strong, especially for old and large organizations, the world of organizations has changed markedly over time. Organizational forms that dominate today differ dramatically from those that held sway a century ago. Chandler (1977) gives a vivid account of the changes in organizational forms in industry over this period. Similar changes can be found in the structures of labor unions, medical care organizations, and government agencies. Thus, changes in social, economic, and political systems apparently do affect organizational structures and practices.

The major gaps in our understanding of organizational change concern the actual dynamics—*exactly how does change in larger systems affect the distribution of organizational forms in society?* In particular, how much of the change in the organizational world comes about through tinkering (adapting organizational strategies and structures) and how much through replacement of one kind of organization by another? We are just beginning to learn about the relative rates of the various processes, which are crucial to answering this question.

Perspectives on Organizational Change

The contemporary literature contains at least three broad perspectives on organizational change. They all emphasize that uncertainty is an inescapable problem for organizations and plays the key role in shaping their structure and action.

The most widespread view, *rational-adaptation theory*, argues that organizational structures are consciously chosen solutions to certain environmental problems. It suggests that the observed variability in the world of organizations reflects planned changes of strategy and structure in response to environmental uncertainties, threats, and opportunities. As a theory of change, this perspective holds that organizations identify threats and opportunities and reshape structures to mitigate threats and exploit opportunities. This approach is mainly directed at explaining the success of large and powerful organizations, those that have managed to adapt well to changing environmental demands.

There are numerous variants of this approach, which differ widely in some ways. Contingency theories stress the need for organizations to design structures that buffer their production activities (the so-called technical core) from uncertain environmental variations (Lawrence and Lorsch, 1967; Thompson, 1967). Thus, optimal organizational design is contingent on the nature of the production process and of environmental variations. When either production processes or the pattern of environmental changes shift, organizations attempt to alter their structures, according to this view. In a similar vein, resource-dependence theory argues that organizations must take action to eliminate sources of uncertainty in the environment (Pfeffer and Salancik, 1978). When sources of uncertainty change, organizations are forced to alter their strategies and structures to resolve new threats to their resource flows.

An institutional approach, discussed at greater length below, holds that organizational structures are rationally adapted to environmental demands, but that the key demands are often normative and symbolic (DiMaggio and Powell, 1983; Meyer and Scott, 1983). In this view, organizations demonstrate their competence within spheres of action and maintain flows of essential resources by displaying appropriate symbols. Such symbolism is often coded in structures. For example, firms display their commitment to planning by creating planning committees or boards of directors and by creating planning departments. What these units actually do is much less important than their mere existence, according to current institutional theories. Moreover, as fads and fashions in organizational designs change, organizations are expected to reshape their structures accordingly. As in the cases of contingency theory and resource-dependence theory, the variability of structures in the world of organizations is assumed to reflect planned adaptations to changing environmental demands.

A second perspective, *random-transformation theory*, claims that organizations change their structures mainly in response to internal politics and other endogenous processes, especially the search for solutions to problems of uncertainty. Because there is much randomness in the character of the search, such changes are only loosely coupled with the desires of organizational leaders and with the demands and threats of environments (March, 1981; March and Olsen, 1976; Weick, 1976).

The third perspective, *ecological-evolutionary theory*, holds that most of the variability in organizational structures comes about through the creation of new organizations and organizational forms and the replacement of old ones (Aldrich, 1979; Carroll, 1984; Freeman, 1982; Hannan and Freeman, 1977; McKelvey, 1982; Nelson and Winter, 1982; Stinchcombe, 1965).

These three perspectives disagree on the sources of organizational diversity. According to rational-adaptation theory, the diversity of organizational forms in society reflects the diversity of environmental problems that must be solved. If the environment becomes more differentiated, diversity will increase; if it becomes less differentiated, diversity will decline. The random-transformation perspective suggests that diversity reflects mainly the peculiar local and random character of problem solving in each organization. Finally, the ecological-evolutionary perspective states that diversity depends on the arrival rate of new organizations and on their diversity, on patterns of environmental variation, and on competitive dynamics within organizational populations and communities.

Progress in explaining organizational diversity and change requires understanding both the nature of organizational change and the degree to which it can be planned and controlled. The remainder of this essay concentrates mainly on the first issue: does most of the observed diversity in organizational features reflect changes in existing organizations, whether planned or not, or does it reflect changes in populations with relatively inert organizations replacing one another? In other words, does change in major features of organizations over time reflect mainly adaptation or selection and replacement?

An Ecological-Evolutionary Approach

If organizations are subject to strong inertial pressures and face changeable, uncertain environments, there are strong parallels between change in organizational populations and change in biotic populations. In this case it may be useful to analyze selection and replacement in *populations* of organizations. As I try to illustrate below, this shift in focus has opened new and interesting questions.

A population perspective concentrates on the sources of variability and homogeneity of organizational forms. It considers the rise of new organizational forms and the demise of existing ones. In doing so, it pays considerable attention to population dynamics, especially the processes of competition among diverse organizations for limited markets.

All accepted theories of biotic evolution share the assumption that innovation, the creation of new strategies and structures, is random with respect to adaptive value. Innovations are not produced because they are useful, they are just produced. If an innovation turns out to have adaptive value, it will be retained and spread through the population with high probability. In this sense, evolution is blind. How can this view be reconciled with the fact that human actors devote so much attention to predicting the future and to developing strategies for coping with expected events?

Most theorists assume that change in organizational populations is Lamarckian, that major changes in the forms of organization come about through learning and imitation. Many kinds of organizations do devote resources to learning and espionage, often seeking to copy the forms of their more successful competitors. In a rough sense, organizations reproduce themselves either by setting up new organizations or by spinning off personnel with the requisite knowledge to copy the form. Nelson and Winter (1982) have developed explicit models of such Lamarckian evolutionary change in populations of business firms.

Another line of theory holds that change in evolutionary populations is more Darwinian than Lamarckian (Aldrich, 1979; Hannan and Freeman, 1977, 1984; McKelvey, 1982). This work argues that inertial pressures prevent most organizations from radically changing their strategies and structures once established. It also argues that only the most concrete features of technique can be easily copied and inserted into ongoing organizations. Finally, it emphasizes density-dependent constraints on adaptation by individual organizations: although it may be in the interests of leaders of many organizations to adopt a certain strategy, the capacity of the system to sustain organizations with that strategy is often quite limited. Only a few can succeed in exploiting such a strategy, and "first-movers" have decided advantages.

Even when actors strive to cope rationally with their environments, action may be random with respect to adaptation as long as the environments are highly uncertain or the connections between means and ends are not well understood. It is the *match* between action and environmental outcomes that must be random on the average for Darwinian selection models to apply. In a world of high uncertainty, adaptive efforts by individuals may turn out to be essentially random with respect to future value.

The realism of Darwinian mechanisms in organizational populations also turns on the degree to which change in organizational structures can be and adapt strategies accordingly, and that organizations simply mirror the intentions of rational leaders. Then organizational adaptations would be largely nonrandom with respect to future states of the environment. On the other hand, if March and others are right, organizational change is largely uncontrolled, and organizations staffed by rational planners may behave essentially randomly with respect to adaptation. In other words, organizational outcomes may be decoupled from individual intentions; organizations may have lives of their own. In this case it is not enough to ask whether individual humans learn and plan rationally for an uncertain future. One must ask whether organizations as collective actors display the same capacities.

The applicability of Darwinian arguments to changes in organizational populations thus depends partly on the tightness of coupling between individual intentions and organizational outcomes. At least two well-known situations generate loose coupling: diversity of interest among members and uncertainty about means-ends connections. When members of organizations have diverse interests, organizational outcomes depend heavily on internal politics, on the balance of power among factions. In such situations collective outcomes cannot easily be matched rationally to changing environments.

When the connections between means and ends are uncertain, carefully designed adaptations may have completely unexpected consequences. Moreover, short-run consequences may often differ greatly from long-run consequences. In such cases, it does not seem realistic to assume a high degree of congruence between designs and outcomes.

Thus, it may be useful in analyzing patterns of long-term change in organizational forms to supplement Larmarckian theories with Darwinian ones. The fact that members of organizations plan rationally for change and that organizations often develop structures designed to plan and implement change does not undercut the value of this view as long as organizations are political coalitions and environmental change tends to be highly uncertain.

Organizational Diversity

An ecological-evolutionary approach directs attention primarily to organizational diversity. It seeks to answer the question: Why are there so many (or too few) kinds of organizations? Addressing this question means specifying both the sources of increasing diversity, such as the creation of new forms, as well as the sources of decreasing diversity, such as competitive exclusion of forms. In other words, an ecology of organizations seeks to understand how social conditions affect the rates at which new organizations and new organizational forms arise, the rates at which individual organizations change structures, and the rates at which organizational populations die out. In addition to focusing on the effects of larger social, economic, and political systems on these rates, an ecology of organizations also emphasizes the dynamics that take place within organizational populations.

Questions about the diversity of organizations in society might seem to be of only academic interest. In fact, these issues bear directly on important social issues. Perhaps the most important is the capacity of a society to respond to uncertain future changes. Organizational diversity within any realm of activity such as medical care, microelectronics production, or scientific research constitutes a repository of solutions to the problem of producing certain sets of collective outcomes. These solutions are embedded in organizational structures and strategies. The key aspects of these solutions are usually subtle and complicated. In any large organization, no single individual understands the full range of activities and their interrelations that constitute the organizational solution. Moreover, the subtle aspects of the structure such as "climate" or "culture" defy attempts at formal engineering specification. Therefore, it will often prove impossible to resurrect a form of organization once it has ceased to operate. If so, reductions in organizational diversity imply losses of organized information about how to adapt (produce) to changing environments.

Having a range of alternative ways to produce certain goods and services is valuable whenever the future is uncertain. A society that retains only a few organizational forms may thrive for a time. But once the environment changes, such a society faces serious problems until existing organizations can be reshaped or new ones created. Since reorganization is costly and may not work at all for the reasons stated above (and because new organizations are fragile), it may take a long time to adapt to the new conditions. A system with greater organizational diversity has a higher probability of having in hand some solution that is satisfactory under changed environmental conditions. Adaptation to changing environments in such cases means mainly reallocating resources from one type of existing organization to another.

The notion that diversity of organizational forms is a useful hedge against uncertain future changes in environments parallels a classic evolutionary argument. It is, for example, the same kind of argument that has been made against going overboard with the so-called Green Revolution in agriculture. The spread of single strains of crops implies a great reduction in genetic diversity, which may prove problematic if new kinds of pests arise to which the "miracle" crops are vulnerable.

Organizational diversity affects society in another way. Since careers are played out in organizations, the distribution of opportunities for individual achievement depends on the distribution of organizational forms. When diversity is high, individuals with different backgrounds, tastes, and skills are more likely to find organizational affiliations that match their own qualities and interests. For example, the fact that virtually every industry in the United States contains a sizable number of small businesses allows ethnic and immigrant communities to create "ethnic enclaves" within which to develop protected career paths. The presence of such niches in the economy, one kind of organizational diversity, has proven crucial to the economic success of at least some ethnic communities.

Diversity is also valued in its own right. Consider the case of the daily press. It is widely agreed in this country that diversity of editorial opinion is a social good and ought not to be sacrificed to economies of business concentration in the industry. Similar views pertain to schooling, higher education, research laboratories, and all sorts of art-producing organizations.

How do social, economic, and political environments affect organizational diversity? Almost all attempts to answer this question focus on the controlling role of uncertainty—stable and certain environments almost surely generate low levels of diversity. The main theoretical question is: How does environmental uncertainty affect diversity?

NICHE THEORY

One line of current research in sociology attempts to explain variations in organizational diversity within the context of what population ecologists call niche theory. The concept of niche is used in population ecology to refer to the set of conditions under which some form of life can perpetuate itself. The niche is a mapping between states of the environment and probabilities of expansion of numbers. Thus the niche summarizes the environmental dependence of a population. Much of the recent progress in bioecological theory has involved embedding naturalistic observations on the structure of niches in nature within Darwinian evolutionary theory (see Roughgarden, 1979).

Much work on the evolution of the niche emphasizes niche width. Some forms, called generalists, persist under a very broad range of environmental conditions. Others, called specialists, thrive only in highly specific environments. Niche theories attempt to explain how patterns of environmental variations affect the evolution of niche widths in biotic communities, that is, how they affect the reproductive success of specialists and generalists.

John Freeman and I have argued that many of the classic problems of environmental uncertainty and organizational structure can be recast profitably as problems of organizational niche width (Freeman and Hannan, 1983; Hannan and Freeman, 1977). Organizations clearly vary on this dimension. In our study of American labor unions, we find the Siderographers Union, which seeks to organize a labor force that numbers in the hundreds (siderographers print currency and stock and bond certificates), and the Teamsters, who try to organize almost anyone who works. Likewise, the economy contains firms that produce a single product and others that produce a great range of products. The connection between niche width and diversity is straightforward. To the extent that social trends favor generalist organizations, organizational diversity will decline. But, if specialist organizations have adaptive advantages, the society will contain many diverse specialists. In other words, the dynamics of organizational niche width constrain organizational diversity.

Theories of organizational niche width deal with a ''jack-of-all-trades, master-of-none'' problem. There are obvious trade-offs between the capacity to withstand a wide range of environmental variations and the capacity for high levels of performance in any one environmental state. In part, these trade-offs concern organizational ''slack'' or excess capacity. The ability to tolerate diverse environments requires the maintenance of numerous routines, patterns of activity that can be invoked by organizational subunits. As Nelson and Winter (1982) convincingly argue, organizations remember by doing, and the capacity to perform a routine declines rapidly with disuse. Generalists, who possess a wide repertoire, must devote considerable resources to simply maintaining the readiness of seldom-used routines. Because generalists must commit so many resources to maintaining and rehearsing routines, they sacrifice efficiency and effectiveness in performing any single routine. Therefore, at least some specialists usually perform better than generalists in any particular environmental state. Whether there are any adaptive advantages to generalism depends on the rate of environmental change and on the patterns of changes.

Levins (1968) proposed a theory of niche width that seems applicable to these organizational questions: Optimal niche width depends on three factors: (1) the magnitude of environmental variations relative to the adaptive capacity of the population, (2) the uncertainty of environmental changes, and (3) the grain of environmental changes. Certainty refers to the odds that the environment will turn up in any particular state; in a maximally uncertain environment each of the possible states is equally likely. Grain refers to the typical durations of environmental states. In fine-grained environments durations are short relative to (unconditional) life expectancy; in coarse-grained environments, typical durations are long. The importance of this distinction is that fine-grained environments ought to be experienced (from a selection perspective) as a weighted average of the states. But populations in coarse-grained environments cannot adapt to some average value; they must have the capacity to withstand long spells in any particular state.

Freeman and I formalized the implications of Levins' model for death rates of specialist and generalist organizations in alternative regimes of environmental variations (Freeman and Hannan, 1983). The implications of this model agree in part with the existing literature but differ in one important respect. The conventional wisdom holds that uncertain environments always favor generalist organizations (see, for example, Katz and Kahn, 1978:131; Lawrence and Lorsch, 1967:8; Pfeffer and Salancik, 1978; and Thompson, 1967:34–37). The model based on niche theory implies that uncertainty favors generalists only in coarse-grained environments. We tested this hypothesis using data on the lifetimes of restaurant firms in 18 California cities. We find that the effect of uncertainty on the relative death rates of specialists and generalists does interact with grain in the predicted way. That is, the niche theory improves on existing theory in explaining the dynamics of niche width in this organizational population.

Carroll (1985) has taken a slightly different approach to studying organizational niche width. He points out that the life chances of specialist and generalist organizations depend on the density of each type in the environment. Imagine a market with a center (high, concentrated demand) and a periphery consisting of pockets of heterogeneous demand. In the absence of competition, all organizations in the market concentrate on the center of the market. When the number of organizations competing in the market is high, the largest and most powerful generalists will typically dominate the center. If generalists are numerous, some of them will be forced to exploit more peripheral segments of the market. Because their size and power may allow them to outcompete specialists in the periphery, the life chances of specialists deteriorate when there are many generalists in the market. If, however, one or a few generalists come to dominate and push the other generalists completely out of the market, the opportunities for specialists to thrive in the periphery rise. Thus, concentration in a market should have the opposite effects on the life chances of specialists and generalists. As a market (or organizational field more generally) concentrates, the death rates of generalists will rise and those of specialists will fall. Analysis of death rates in populations of local newspaper firms supports this argument (Carroll, 1985).

Our research group is conducting additional research on the dynamics of organizational niche width among labor unions and semiconductor manufacturing firms. A number of other groups are working on similar issues using different kinds of organizational populations.

INSTITUTIONAL ISOMORPHISM

Organizational ecology also speaks to issues of structural isomorphism—the processes by which organizational structures become matched to features of the environment. Indeed, the initial work on the population ecology of organizations was stimulated partly by Hawley's (1950, 1968) classic argument that organizational structures become structurally isomorphic to those organizations that control flows of resources into a local system. An example of this process is the spread through research universities of planning and budget offices whose internal arrangements are close copies of those of the federal agencies from which universities obtain funding.

Hawley's theory was silent on the processes that generate structural isomorphism. Hannan and Freeman (1977) argued that one route to isomorphism is competition and selection at the population level—competitive isomorphism, as DiMaggio and Powell (1983) call it. But clearly this is not the only one.

One important new line of sociological theory about organizations identifies institutional processes that produce isomorphism. John Meyer and his collaborator (Meyer, 1978; Meyer and Scott, 1983) have argued that many features of organizational structure are symbols for competencies that may or may not exist. The important adaptive problem for organizations, especially those producing products whose quality is hard to measure, is to evoke the appropriate symbols of competence. Moreover, organization builders are constrained by norms of rationality that dictate a limited number of routines and structures.

As general societal processes of rationalization and state expansion proceed, the set of available and endorsed building blocks becomes increasingly homogeneous, and organizational diversity declines. DiMaggio and Powell (1983:147–148) argue as follows:

> Bureaucratization and other forms of organizational change occur as the result of processes that make organizations more similar without necessarily making them more efficient . . . highly structured organizational fields provide a context in which individual efforts to deal rationally with uncertainty and constraint often lead, in the aggregate, to homogeneity in structure,.culture, and output. . . . In the initial stages of their life cycle, organizational fields display considerable diversity in approach and form. Once a field becomes well established, however, there is an inexorable push towards homogenization.

The argument that general norms of rationality and specific organizational agents (like the state, business schools, and professional associations) create pressures for structural homogeneity, and that these pressures are growing in strength is an important one. But there is a countertendency that must also be considered.

Assume that the population of individuals who demand and use se
of organizations is heterogeneous. Assume further that organizational
partly determine the character of organizational outputs. If, as the i
tionalists claim, organizations are becoming more homogeneous, th
tion of demand for organizational outputs that is either unfilled or dissa
with current services will increase. This means that the gains from cr
"deviant" organizations to fill this demand will grow. This situation c
opportunities for "outlaw" entrepreneurs to experiment with new c
zational forms. If any of the experiments are successful, the new
ought to grow rapidly, lowering the homogeneity of the organiza
population.

It seems that industrial breakthroughs are often made outside i
tional channels. The industrial giants in any one era often lack the
sight and flexibility to exploit radically new technologies and strat
Sometimes new modes of production are inconsistent with standa
erating procedures in the industry, producing conflict that drives cr
individuals out of giant firms into entrepreneurial ventures. The
semiconductor industry provides an instructive example (Brittai
Freeman, 1980). Each of the large vacuum tube producers tried its
at the production and marketing of semiconductor devices and f
The industry became dominated by newly created firms. Almost 30
into the history of the industry, there is still very rapid turnover
list of leading firms.

The crucial organizational innovations that create new industrie
new goods and services arise mainly outside the highly institution
sphere. In fact, high levels of institutionalization may be a serio
pediment to innovation. Understanding the forces that create org
tional diversity requires analysis of the social forces that shape att
to create new forms of organizations and of the selection processe
apply to new forms.

DISCUSSION

An ecological-evolutionary perspective on organizational chang
onciles the major insights of the two classic traditions of organiz
theory and research. It assumes along with rational-systems perspe
that organizations are designed as tools to achieve collective ends. B
organizations compete among themselves for scarce resources, membe
and legitimacy, efficiency in mobilizing each of these affects survival ch
In this sense, organizations face efficiency tests. However, the effi
testing assumed in current ecological theory is much more complicate
simple testing for technical efficiency in producing some product or se

Efficiency in mobilizing resources or in currying political favor may often be more decisive in affecting survival chances than narrow technical efficiency. Thus the "rationality" involved in organizational selection processes may be considerably broader than that envisioned by Weber and Taylor. Nonetheless, ecological-evolutionary perspectives pay explicit attention to efficiency testing, broadly defined.

Ecological-evolutionary perspectives also build on the natural-systems notion that organizations take on lives of their own. Because organizations must delegate decisions to human actors, they cannot escape processes of political conflict and the creation of subgroup norms and loyalties. Initial patterns of action tend to become enduring bases of political bargaining and of group loyalties. Subsequent attempts to change drastically the structure of an organization encounter both self-interested political objections from subgroups who will lose resources and normative objections to changing rules and structures that have become infused with symbolic value. For these and other reasons, organizations seldom function exactly as planned and are very difficult to reshape.

Inertia and change in the composition of organizational populations over time are easy to reconcile within a population perspective. New organizations enter many organizational populations at a reasonably high rate. These new entrants are often the carriers of new strategies and structures, the main source of diversity of forms. At the same time, existing organizations drop from sight either by simply disbanding or merging with other organizations. The merger process seems to have become the main vehicle by which large and powerful organizations cease to have independent effects on the society. Thus, the populations of business firms, government agencies, and others have changed over time in response to the creation of new firms and agencies carrying new forms and following new agendas. They have also changed in response to mergers among existing firms and agencies.

This view of organizational change directs attention to social policies that affect the rate of creation of new organizations. It suggests that discussion of industrial policy, for example, should pay less attention to established, giant firms than to the social, political, and economic processes that affect the rate at which new firms are started and the life chances of new firms using innovative strategies and structures. More generally, it points to the importance of organizational diversity to society and emphasizes the need to better understand how social policies affect such diversity.

* * *

Many of the ideas discussed here were worked out in collaboration with John Freeman. Susan Olzak made helpful comments on an earlier draft.

REFERENCES

Aldrich, Howard E.

1979 *Organizations and Environments*. Englewood Cliffs, N.J.: Prentice-Hall.

Arrow, Kenneth J.

1974 *The Limits of Organization*. New York: W.W. Norton.

Blau, Peter M.

1955 *The Dynamics of Bureaucracy*. Chicago: University of Chicago Press.

Brittain, Jack, and John Freeman

1980 Organizational proliferation and density-dependent selection. In John Kimb Robert Miles, eds., *Organizational Life Cycles*. San Francisco: Jossey-Bass.

Carroll, Glenn R.

1984 Organizational ecology. *Annual Review of Sociology* 10:71–93.

1985 Concentration and specialization: dynamics of niche width in populations of zations. *American Journal of Sociology* 90:1262–1281.

Chandler, Alfred D., Jr.

1977 *The Visible Hand: The Managerial Revolution in American Business*. Car Mass.: Belknap.

Coleman, James S.

1982 *The Asymmetric Society*. Syracuse, N.Y.: Syracuse University Press.

Cyert, R.M., and J.G. March

1963 *A Behavioral Theory of the Firm*. Englewood Cliffs, N.J.: Prentice-Hall.

DiMaggio, Paul J., and Walter W. Powell

1983 The iron cage revisited: institutional isomorphism and collective rationality nizational fields. *American Sociological Review* 48:147–160.

Freeman, John

1982 Organizational life cycles and natural selection processes. In Barry M. S Lawrence L. Cummings, eds., *Research in Organizational Behavior*, Vol. 4 wich, Conn.: JAI Press.

Freeman, John, and Michael T. Hannan

1975 Growth and decline processes in organizations. *American Sociological Review* 228.

1983 Niche width and the dynamics of organizational populations. *American Jo Sociology* 88:1116–1145.

Hannan, Michael T., and John Freeman

1977 The population ecology of organizations. *American Journal of Sociology* 82:9

1978 Internal politics of growth and decline. In Marshall Meyer and others, eds., *ments and Organizations*. San Francisco: Jossey-Bass.

1984 Structural inertia and organizational change. *American Sociological Review* 164.

Hawley, A.

1950 *Human Ecology: A Theory of Community Structure*. New York: Ronald.

1968 Human ecology. Pp. 328–337 in *International Encyclopedia of the Social* New York: Macmillan.

Katz, Daniel, and Robert L. Kahn

1978 *Social Psychology of Organizations*. 2nd ed. New York: Wiley.

Lawrence, Paul, and Jay Lorsch

1967 *Organization and Environment: Managing Differentiation and Integration*. Ca Harvard University Press.

Levins, Richard

1968 *Evolution in Changing Environments*. Princeton, N.J.: Princeton University Press.

March, James G.

1981 Footnotes on organizational change. *Administrative Science Quarterly* 26:563–597.

March, James G., and Johan P. Olsen

1976 *Ambiguity and Choice in Organizations*. Bergen, Norway: Universitetsforlaget.

McKelvey, Bill

1982 *Organizational Systematics*. Berkeley: University of California Press.

Meyer, John W.

1978 Strategies for further research. In Marshall Meyer et al., eds., *Environments and Organizations*. San Francisco: Jossey-Bass.

Meyer, John W., and W. Richard Scott

1983 *Organizational Environments: Ritual and Rationality*. Beverly Hills, Calif.: Sage.

Michels, Robert

1962 *Political Parties: A Sociological Study of the Oligarchical Tendencies of Modern Democracy*. (Translated by E. Pane and C. Pane). New York: Dover.

Nelson, Richard R., and Sidney G. Winter

1982 *An Evolutionary Theory of Economic Change*. Cambridge, Mass.: Belknap.

Ogburn, William F.

1933 The influence of invention and discovery. In the President's Research Committee on Social Trends, *Recent Social Trends in the United States*. New York: McGraw-Hill.

Parsons, Talcott

1960 *Structure and Process in Modern Societies*. Glencoe, Ill.: Free Press.

Perrow, Charles

1979 *Complex Organizations: A Critical Essay*. 2nd ed. Glenview, Ill.: Scott Foresman.

Pfeffer, Jeffrey

1981 *Power in Organizations*. Marshfield, Mass.: Pitman.

Pfeffer, Jeffrey, and Gerald Salancik

1978 *The External Control of Organizations: A Resource Dependence Perspective*. New York: Harper and Row.

President's Research Committee on Social Trends

1933 *Recent Social Trends in the United States*. New York: McGraw-Hill.

Roethlisberger, F.J., and William J. Dickson

1939 *Management and the Worker*. Cambridge, Mass.: Harvard University Press.

Roughgarden, Jonathan

1979 *Theory of Population Genetics and Evolutionary Ecology: An Introduction*. New York: Macmillan.

Scott, W. Richard

1981 *Organizations: Rational, Natural, and Open Systems*. Englewood Cliffs, N.J.: Prentice-Hall.

Selznick, Philip

1948 Foundations of the theory of organizations. *American Sociological Review* 13:24–35.

1949 *TVA and the Grass Roots*. Berkeley: University of California Press.

Stinchcombe, Arthur S.

1965 Social structure and organizations. In James G. March, ed., *Handbook of Organizations*. Chicago: Rand McNally.

Thompson, James D.

1967 *Organizations in Action*. New York: McGraw-Hill.

Tilly, Charles

1978 *From Mobilization to Revolution*. Redding, Mass.: Addison-Wesley.

Weber, Max

1978 *Economy and Society: An Outline of Interpretive Sociology*. 2 vols. Berkeley: University of California Press.

Weick, Karl

1976 Educational organizations as loosely coupled systems. *Administrative Science Quarterly* 21:1–19.

Williamson, Oliver E.

1975 *Markets and Hierarchies*. New York: Free Press.

Macroeconomic Modeling and Forecasting

LAWRENCE R. KLEIN

ORIGINS OF THE SUBJECT

Historical research may uncover some obscure roots or primitive example of macroeconomic models, estimated from numerical records and meant to be realistic, but I doubt that anything clearly in the spirit of the present effort can be found prior to the investigations of J. Tinbergen, first for his own country, the Netherlands, in the early 1930s, and later for the League of Nations' analysis of the economy of the United States, intended for a better understanding of the Great Depression (Tinbergen, 1939).

Tinbergen's contribution was to show that the essence of the macro economy could be expressed compactly in an equation system, that these systems could be fit to real-world data, and that revealing properties of the economy could be analyzed from such systems. To a large extent, Tinbergen was interested in the cyclical properties of the system. That was his main reference point in studying the American economy in the context of the stock market boom of the 1920s, followed by the Great Crash and recovery during the Great Depression of the 1930s. He was plainly impressed and inspired by the implications of the Keynesian Revolution, but his greatest work on econometric models, that of the United States for the League of Nations, was never put to practical use in the national scene. His model estimation for the Netherlands formed the basis for Dutch postwar economic policy implementation at the Central Planning Bureau, which he directed after World War II.

There was a hiatus, naturally, caused by the war, but during the closing months of 1944, J. Marschak assembled a research team at the Cowles

Commission of the University of Chicago. The Cowles Commission was founded and supported by Alfred Cowles, originally for the study of security prices and eventually for the investigation of mathematical and statistical methods in economics. I was recruited for the Cowles team for the express purpose of taking up the thread of work initiated by Tinbergen. Several lines of thought converged at the Cowles Commission during the middle and late 1940s. These were:

- The concept of a mathematical model of the macro economy
- An emerging theory of econometric method
- A growing body of statistical data on the national economy

The macro model concept built on the intense intellectual discussions among economists about the interpretation of the theories of J.M. Keynes in *The General Theory of Interest Employment and Money*. The mathematical formulations of that theory by J.R. Hicks (1937) and O. Lange (1938) formed the basis for a whole new way of thinking about the aggregative economy. F. Modigliani had written a provocative article extending the Keynesian analysis (Modigliani, 1944), and I had just completed the dissertation draft of *The Keynesian Revolution* in 1944. The mathematical models in these writings lent themselves well to manipulation to study the movement of principal magnitudes of the economy and were formulated in terms of measurable concepts. It was essentially a case of ''actors in search of a play.'' The Keynesian theory was simply crying for econometric implementation.

In a miniature, condensed, and aggregative sense, the Keynesian theory was a simultaneous equation version of the Walrasian system for the economy as a whole. In the nineteenth century, L. Walras, professor at Lausanne, formulated a view of the economy as a result of the solution of a set of simultaneous equations. His set referred to the detailed micro economy and, conceptually, to an enormous system of *n* equations in *n* unknowns, where *n* is as large as the goods and services of an economy are numerous, i.e., in the millions, or billions, or even larger. At a macroeconomic level the Keynesian economic theory recognized the simultaneity clearly. For example, it was noted that aggregate income affected aggregate spending in the economy, while at the same time aggregate spending affected the generation of aggregate income. But statistical practice did not take this simultaneity properly into account. This idea was exploited by T. Haavelmo, much inspired by A. Wald, who contributed a great deal of the statistical thinking—as far as probability and the laws of inference are concerned—and also the dynamics. Two important papers were produced that shaped the statistical approach of the Cowles Commission team (Haavelmo, 1943; Mann and Wald, 1943). This approach has not flourished as much as the approach of building macroeconometric models for practical and theoretical application, but it was instrumental in providing a deep understanding of econometric method. The statistical approach was a moving force in the formative days, even though it is not preeminent at the present time.

The actors and the play came together in the actual statistical calculation of models. They began to emerge as early as 1946 and were first used by the Committee for Economic Development in assessing economic prospects for the postwar world. The emphasis was different from Tinbergen's. The principal goal was to build models in the image of the national income and product accounts for the purpose of making forecasts and for guiding economic policy. The kinds of policy formulations implicit in Keynes's *General Theory* were plainly operative in these systems.

A PERIOD OF EXPANSION

The history of the development of models during this period and in the subsequent two decades or more is traced by Martin Greenberger et al. (1976) and Ronald Bodkin et al. (1980). It consists of tracing the models from the Cowles Commission at Chicago, to the University of Michigan (Ann Arbor), Canada, the United Kingdom, and elsewhere up to the present day, when there are literally hundreds in daily operation all over the world. The major actors in this history were, in addition to the author, Colin Clark, Daniel Suits, Arthur Goldberger, R.J. Ball, Otto Eckstein, J. Duesenberry, and Gary Fromm (Clark, 1949; Klein and Goldberger, 1955; Suits, 1962; Klein et al., 1961; Duesenberry et al., 1960).

During the period of enthusiastic development at the Cowles Commission it was thought that applications of the most sophisticated and powerful methods of statistical analysis would provide a breakthrough in practical accomplishments, but complexity of computation remained a bottleneck. Only demonstration models could be given a full statistical treatment, and that was very laborious.

The use of cross-section data (surveys of individual economic units—households and establishments), and finer units of observation in time series (quarterly and monthly data), were looked to as other routes for a breakthrough. Cross-section data were "noisy" although revealing. Monthly and quarterly data were highly serially correlated but helpful in cyclical analysis. Trend data, in the form of decade averages over long periods of time, were also examined for possible leads, but they were too smooth to make an enormous difference.

In the early 1960s a breakthrough did occur, in the form of the electronic computer, which was harnessed to the needs of econometrics. In the 1950s there were some early attempts at massive computer use, which worked well for selected aspects of econometric computation; but it was only through the use of the computer in successive stages that very significant achievements were realized. It is noteworthy that at a 1972 conference in honor of John von Neumann's contributions to the development of the electronic computer, held at the Institute for Advanced Study in Princeton, most of the reports claimed that the computer was of only moderate significance in advancing scholarly subjects in the sciences. Some of the contributions bordered on the cynical side. But economics was an exception. In my own paper, I claimed that the computer absolutely transformed research in quantitative economics and in econometric model building in particular.

The use of cross-section and sample survey observations in general has a long history in econometrics, both theoretical and applied. These observations were used more for the study of building blocks at the micro level of analysis, but as the computer became more available for econometric research, it became more plausible to build models that evolved into complete systems from the micro units. These are micro simulation models, introduced at an early stage by Guy Orcutt (Orcutt et al., 1961). Large segments of the total economy, if not the entire macro economy, can be modeled in this way. Such systems are not as widespread in use as are aggregative models of the economy as a whole, but progress in provision of data, techniques of data management, and system computation will enable model building to go in the direction of micro simulation systems in the future. At the very least, they will be revealing for the study of particular aspects of human behavior patterns because they get right down to the basic agents of behavior.

By 1964 it was possible to compute full-information maximum likelihood estimates of equation systems with 20 or more equations. The usual examples had dealt with systems of 3, 4, or 5 equations using hand methods. The computational problem, as far as statistical method in econometrics was concerned, was fully resolved during the 1960s. Programs for estimation in nonlinear systems, autoregressive correction, estimation of distributed lags, ridge regression, generalized regression, and many other estimation problems were made available for routine use. All the bottlenecks that had appeared earlier were suddenly broken. Econometric scholars were able to handle data much better and explore data much more extensively in the search for good estimates, but there was no seeming increase in accuracy, efficiency, or applicability of econometric models from this particular line of research. The next real breakthrough came in connection with some research on the Brookings model for the macro economy of the United States. This was a quarterly model put together by a team of specialists working under the auspices of the Committee on Economic Stability of the Social Science Research Council. The work started in 1960, but a model was not available until 1963 and was then transferred to the Brookings Institution.

The principal problem was the computation of a solution to a system of some 300 equations, which seemed very large at the time. After much detailed experimentation, a method of solution was found in 1965. It was a form of a well-known Gauss-Seidel algorithm, which proceeds by iterative substitution of partial solutions from one equation to the next in a long succession through a whole system. This method was tedious but inexpensive and fast on an electronic computer. Although it involved a very large number of calculations, it was accurate, efficient, and workable. It is now a routine method used worldwide for solving systems of simultaneous equations in economics. Once the method had been streamlined for either nonlinear or linear econometric models, the technique of simulation was extensively developed for the analysis of numerical models.

Use of this instrument was a breakthrough in the following senses:

1. Systems of any manageable size could be handled. A system of 100 equations was considered modest (for the first time), and systems of thousands of equations are frequently used. Manageability is governed by the capability of human operators to collect data, have it ready for use, and understand the working of the system.
2. Economic concepts such as multipliers became generalized into alternative policy simulations, scenarios, dynamic or static solutions, or stochastic simulations. All of these enabled workers in the field to do much more with models, to understand their dynamics and sensitivities. For policy application in both private and public sectors, extensive simulation analysis is essential.
3. Frequency response characteristics of dynamic systems could be studied. Eigenvalues in linear or linearized systems could be studied; methods of optimal control could be applied.
4. The presentation of econometric results for the noneconometrician or even the noneconomist became possible. Solutions of abstract mathematical systems could be presented in the form of instructive accounting tables, graphical displays, time patterns, and condensed reductions of large, complicated systems.
5. Error statistics could be computed. With stochastic simulation methods probability limits on forecast error bands or regions could be evaluated. Extensive recordkeeping for model performance became possible.

The forms of calculation and analysis just listed were always possible. In linear systems they could be expressed in closed form relationships for most cases, but they could never have been done on a significant scale and would have attracted only very few patient individuals to the field of applied econometrics. Those of us who toiled for 2 or 3 days to make one alternative analysis of a 20-equation system still marvel at the hourly accomplishments of the computer with a model. We could never have dreamed, in 1945, that we would be using models so intensively, so extensively, with such an audience, or with as much precision as is realized today. The computer is the key.

At the present time econometric software takes one from beginning to end (''cradle to grave''). Data come in machine-readable form from the primary, usually official, source. The data are managed by software designed to take out seasonal variation, arrange in order for computation, correct for inflation, and form transformations (ratios, sums, logarithms, etc.). Estimation routines are put to work directly from updated data files. The estimated equations are screened and selected for use in models. Simulation routines arrange the equations for dynamic solution in various forms (stochastic, deterministic, steady state, short run). All these things are done internally—within the computer. Finally, tables, charts, and graphs are generated as reports. This is the patterned sequence for the present computer age.

Every breakthrough has its drawbacks. Today's econometrician can make horrendous mistakes because the computer stands between the original material and the final results. Modern investigators do not look at each sample value with the same care that older econometricians used when the data were personally handled. A certain number of initial mistakes or nonsensical results have to be tolerated as payment for the enormous amount of good material that is generated. Only the ''old hands'' know where to look for troubles and avoid pitfalls, because the typical modern investigator does not want to search into or behind the computer processing unit.

CONTRIBUTION TO THOUGHT

The computer is a facilitator but it does not guarantee good analysis or usage, enabling us to produce econometric findings on a large scale, presentable in convenient form. It is now time to consider the impact of this effort, not particularly from the viewpoint of the immediate users, but from the viewpoint of scholarly thought.

Econometric methods are often used to test economic theory. The models are themselves often, and preferably, based on received economic theory. Whether or not they fit the facts of economic life should tell us something about the validity of the underlying theory.

Some direct tests of economic theory have been decisive, but we often come up against the fact that the data of economics, which form the sampling basis for econometric inference, are not sharp enough or abundant enough to come to decisive conclusions in many cases. More than one hypothesis is often consistent with a given body of data.

We have, however, been able to reject the crudest and most simplistic of theories. The data, and tests based on these data, have been conclusive in rejecting crude acceleration principles or the crude quantity theory of money:

$$I_t = a\dot{C}_t + e_t$$
$$M_t = k\,[\mathrm{GNP}\ (\$)]_t + u_t$$

where

I_t = net real investment in period t

$\dot{C}_t$ = rate of change of real consumption during period t

M_t = money supply at time t

$[\mathrm{GNP}\ (\$)]_t$ = nominal GNP during period t

e_t, u_t = random errors

a, k = parameters

If we hypothesize that I_t is proportional to $\dot{C}_t$ apart from an additive random error that is obtained from a fixed distribution with finite variance and no serial correlation, we would find that the data do not support this model.

It is, of course, a simple model, and if it is generalized by replacing $\dot{C}_t$ by total real output and introducing other variables such as capital cost and capital stock we get the generalized accelerator, which does appear to fit the facts. That does not mean that we have proved that the generalized accelerator is the *only* investment function that fits the facts; indeed, there are others that are consistent with observed data, but we have been able to reject the crude, and original, version of the accelerator hypothesis.

The same is true of the crude quantity theory. Either Milton Friedman's generalization, by introducing distributed lags in the price and real output factors (components) of [GNP ($)], or some extended liquidity preference version of the Keynesian theory both fit the facts. We cannot discriminate definitively between the monetarist hypothesis of Friedman or the portfolio hypothesis of Keynes, yet we have been able to reject the crude form of the theory.

There are many examples like these, but it can legitimately be argued that such testing does not take us very far because it leaves an embarrassingly large number of competitive hypotheses still standing as possible explanations of reality. When we go beyond the simple single relationship, however, to fully determined models of the economy as a whole, we find that it is not easy to put together an entire set of relationships that explains, to an acceptable degree, the evolution of the macro economy. It is, admittedly, difficult to settle on a single set of composite relationships for the economy as a whole, yet it is possible to find more than one that passes the conventional statistical tests.

The econometric approach should be used, therefore, painstakingly and slowly, to sift through this multiplicity of systems in replicated applications and to test them by the strength of their predictive powers. It is usually the case that particular models will work well on one occasion or another, but it is not at all easy to find a system that will stand up to predictive testing period after period. If an estimated model does predict well in repeated applications, then we get some evidence about the validity of the hypothesis on which it is based.

Some macroeconometric models that were thought to rest firmly on accepted hypotheses, such as the St. Louis model, the Fair model, or various supply-side models, did so poorly in predictive testing that they were deemed failures (McNees, 1973; Fair, 1976; Andersen and Carlson, 1976).[1] Monetarist economists were enthusiastic about their hypotheses in the late 1960s, but when the St. Louis model, which was based on those hypotheses, came up against the oil embargo, OPEC pricing, food price inflation, and the automobile strike of 1969, its operators declared that it was not meant for short-run prediction of the economy. This statement contradicted their original hypotheses. Eventually the monetarists contended that the model could be used for long-term but not for short-term analysis. In this case, it appeared to fit the data of the economy for awhile, but with repeated use, observations emerged that were not consistent with its underlying theory. The same is true of the original Fair model.

As for the supply-side models that were introduced in the late 1970s, they were never tested against the facts, but when they were estimated and confronted with the data of 1981–1982 they failed to stand up as maintained hypotheses. More conventional models correctly predicted that this would be the outcome.

Given only limited success for macroeconometric model analysis in testing economic theory, how has it otherwise contributed to economic thought? The main contributions have been to decisionmakers, the users of econometric information in the public and private sectors. They are usually administrators and executives in large organizations and enterprises. Legislators also fall into this group. In a sense, the utility of these models is indicated by the degree of their use. There are systematic records of their forecast

[1] The use of supply-side models is quite new; fully documented citations will not be ready for a few more years.

history but no systematic records of their performance in decisionmaking. Since only one outcome is actually observed, it is impossible to judge their accuracy with respect to unobserved alternatives that tend to be considered by the decisionmaker in reaching a choice.

Decisionmakers say that they make better choices than they would otherwise be making without the use of such models, and that econometric models are the only tool available in a large variety of situations. This is why econometric model-building activity is expanding so rapidly all over the world.

To a large extent, the first models were national in scope and fitted together with the emerging body of data on national income and product. But many subnational and supranational macroeconometric models, for industries, markets, or regions of a nation, are now either in use or being built. Many of these are designed in integrated feedback mode with comprehensive macro models, while many are also built in satellite mode, without feedback.

At the international level we now have many models connecting different countries and regions in a macro world system. One of the first of such systems was designed as an international trade model by J.J. Polak (1953).[2]

In 1969 Project LINK was initiated, with the objective of consistently tying together models of the main trading nations to analyze the international transmission mechanism (Ball, 1973; Klein et al., 1982). The project now consists of the interrelated network of models for 25 industrial countries—the Organisation for Economic Co-operation and Development—(OECD), 8 centrally planned economies, and 4 regions of aggregative developing country models. Work is under way to add more than 25 models of individual developing countries. Project LINK is approaching, in the true sense of the word, the status of a world model.

After the implementation of LINK in analyzing the world economy for such things as oil price shocks and predictions of various global totals, many other supranational systems have been designed: INTERLINK, by the OECD; the TSUKUBA-FAIS Model by Tsukuba University, Japan; the Multicountry Model of the Federal Reserve Board; the World Economic Model of the Economic Planning Agency, Japan; the FUGI Model of Soka University, Japan; and the World Model of Wharton Econometrics. These systems vary according to size and focus. Some concentrate on exchange rates and balance of payments flows; others are principally concerned with trade. Some emphasize the long term; others the short term. Nevertheless, they all have a common interest in the global outcome of economic activity

[2] Other early models were COMET and DESMOS (see Barten, 1981; Dramais, 1981).

and will be used with increasing frequency in an economic world that is becoming increasingly interdependent. Most of these systems were developed during the past 10 years, and it is evident that many more will be developed in the period ahead.

Specifically, the systems are used to study:

- the exchange rate system
- trade liberalization or protection
- world business cycle analysis
- worldwide disturbances (oil, food, raw materials)
- international debt problems
- policy coordination among countries
- transfers of capital
- international migration

As new, equally pressing, issues arise the models will be adapted to their analysis.

SOME NEW LINES OF DEVELOPMENT

Econometric model building in the computer age has moved in the direction of the large-scale (1,000-or-more-equation) system with many sectors, rich dynamics, nonlinearities, and explicit stochastic structure. It has never been viewed as an issue of "bigger is better"; it is mainly an issue of detail. In large, detailed systems of this sort a main interest has been the development of scenario analysis. This procedure generalizes the entire concept of the multiplier, which is meant to show the relationship between any particular endogenous variable (y_{it}) and any corresponding exogenous variable (x_{jt}):

$$\frac{\partial y_{it}}{\partial x_{jt}} \qquad \text{other x's unchanged}$$

This general expression includes the original Keynesian multiplier

$$\frac{dGNP}{dI} = \frac{1}{1 - mpc}$$

mpc = marginal propensity to consume
GNP = real gross national product
I = real investment (exogenous)

This simple multiplier expression is designed to show the GNP that would be generated by an increment in fixed investment. For most countries the GNP gain would outstrip incremental investment, making the multiplier metric models of the United States, indicates that the multiplier, after taking a much more elaborate system structure into account, is about 2.0 after a period of about two years' sustained stimulus to investment.

The scenario, as distinct from the multiplier, simply imposes changes on a model. These changes can be at any place and any time; they may alter any element or group of elements. Computer simulation enables the investigator to compare a baseline or reference solution to a model system with a scenario solution. Scenarios can be quite creative—the scenario of disarmament, of harvest failure, of embargo, of stimulative policy, of technical progress. Investigation is virtually unlimited. It is important to have a credible and fully understood baseline solution; the scenario then produces a discrepancy that reflects the investigator's creative inputs. This can be quite revealing and is of the greatest importance for policymakers, decisionmakers, or interested economists. There is unusual interest in a model's forecasts, but there is as much interest in scenario analysis. Scenario analysis permits rapid response to changing situations, such as abrupt shifts in policy, embargoes, strikes, and natural disasters. It is also the natural tool for planning.

It is evident that the rapid, frequent, and flexible implementation of scenarios would not have been possible without the use of the computer. In addition, the report-presentation capabilities of the computer enable the model operator to communicate final results with a high degree of expositional clarity. These applications on a large scale have been available for only about 10 or 15 years. They are undergoing further development, particularly for use with microprocessors. From the point of view of development of thought, this aspect is likely to be mainly of pedagogical significance.

But the use of scenario analysis in the formulation of economic policy is leading in another direction that does have some methodological and theoretical significance for econometrics. The formal theory of economic policy was introduced some 30 to 40 years ago by J. Tinbergen and others. Tinbergen drew a distinction between *targets* and *instruments* of policy. The former are the subgroup of endogenous variables that policymakers want to have at specified values some time in their planning horizon, while the latter are the items of control among the exogenous variables that policymakers can influence. For example, bank reserves are *instruments* affected by the Federal Reserve system's open market operations that are fixed at certain values in order to achieve policy *targets* such as various money supply aggregates. The formal theory establishes the choice of instruments in relation to target goals in the framework of a loss or gain function that the policymakers attempt to optimize, subject to the constraints of the working of the economy, represented by a macroeconometric model. We are not, by a long shot, near the point at which policy can be routinized through this optimization process. We are not able to do what scientists and engineers accomplish with applications of optimal control theory to the operation of physical systems such as a boiler. We have borrowed many useful ideas from the control theory literature, and the development of this stream of econometric analysis is providing a deeper understanding of the workings of models, especially inputs for exogenous variables over distant horizons. These inputs are those that keep the system close to some *a priori* targets.

Control theory calculations may become difficult if used in large-scale systems. This has been the case, particularly, in large international systems with multimodel components, as in Project LINK. For this reason simplified and small systems are frequently used to facilitate the control theory applications to econometrics. This use is not meant to bring control theory to the needs of actual policy application. Eventually, the cumulation of knowledge and even further computational advances will make control theory methods more suitable for use in large, state-of-the-art models.

Despite the advances in econometric model building and the growth in its use, there are skeptics among professional economists. Some argue that models are not sufficiently accurate although users do seem to appreciate their accuracy to the extent that they find them important elements in their own tool kits. They will undoubtedly continue to use macroeconometric models unless and until something better, cheaper, and more convenient is made available.

For some time, analysts who work with different systems have made claims of either superiority, equality, or cheapness, but they have not produced convincing evidence to support their claims. In some respects, time-series models, which are based purely on sample data with no (or minimal) underlying economic theory, claim to be an alternative. At present, it may be said that time-series models, in single equations or systems of equations, produce forecasts that are at least as good as macroeconometric models for short time horizons, say up to six months or less. Time series models do not provide a flexible vehicle for scenario analysis, but they do provide forecasts.

The contest between time series and macroeconometric models will continue on its present course, and it is plausible to believe that each has something to learn from the other, but there is also another possible route in which they may be developed together for construction of a better system.

It is first necessary to digress for the explanation of a controversial issue in connection with application of large-scale models. After a model is estimated from a typical sample of economic data, it is invariably found iat extrapolations outside the sample lose accuracy quickly. For one year or vo, a freshly estimated model may stay close to economic reality, but it has ever proved possible to estimate a model from a time-series sample and use iat model in extrapolation in an automatic way with zero (mean) values ssigned to the stochastic error terms. All such attempts have failed to pass orecasting tests outside the sample. The reasons for this failure are:

1. data revisions between the estimation period, within the sample, and ie extrapolation period;
2. legal or institutional changes in the functioning of the economy;
3. the occurrence of an unusual disturbance (war, strike, embargo, natral disaster);
4. temporary behavior drifts.

Ine way of dealing with some of these problems in small systems is to eestimate the model every period (as often as every quarter) before extraplation. This is entirely possible in small systems but not in the large models resently in use. Instead, common practice is to estimate nonzero values for ie stochastic component to bring the model solution close to reality for the iitial period before extrapolation. In general, this period will be one outside f the sample. We line up the model so that it starts an extrapolation period t the observed values. Its reaction properties are left unchanged unless there as been a statutory or other known structural change.

The adjustments are made to equations on a nonmechanical basis, and ie criticism of model builders' practices is that they are changing their ystems prior to application by a method that is not purely objective. (It iay be better to say, "by a method that is not purely automatic," because bjective criteria *are* used in choosing the nonzero values for the stochastic erms.)

A suggestion by Clive Granger, who is an exponent of time-series methds, may indicate a fruitful alternative that *is* automatic. Granger suggests iat forecasts be made from linear or other combinations of models in order o spread risk in an uncertain exercise. A different combination is the ollowing: Time-series equations can be estimated for each endogenous ariable of a model. These can be automatically recalculated on an up-to-ate basis as new observations are made available. Error values in an xtrapolation period can be assigned to each equation of a model so that ie time-series estimate of the normalized endogenous variable for each quation is obtained.[3] In other words, the model is automatically and obectively adjusted to "hit" the time-series estimate of the value of each

[3] A normalized variable is the variable that carries a unit coefficient in each stochastic equation.

economy by using daily, weekly, or monthly advance estimates of key magnitudes. A strong current-quarter model may provide better estimates of initial values of endogenous variables than can a pure time-series model.

This area of research for improving the adjustment procedure is one that is presently receiving attention and appears to be promising.

Other lines of development are in variable parameter models and in generalized models that transcend the narrow scope of purely economic relationships.

Parameters may vary over time, systematically or randomly. Methods for dealing with these models have been proposed from time to time. There is no immediate breakthrough in sight, but it is an area that merits much attention.

As for enlarging the scope of models, we have attempted to bring in more policy variables and introduce political reactions. This is particularly the case for intervention in the foreign exchange markets in this era of floating rates.

Economists tend to draw a line and delegate responsibility to demographers for birth, death, morbidity, and other population statistics; to criminologists for crime statistics; to psychologists for attitudinal variables; to political scientists for voting magnitudes, and so on. Usually, these external magnitudes are classified as exogenous variables, but many interact with the economic system in feedback relationships. Over the years the scope of endogeneity has expanded. Many models generate age-sex-race distributions of labor force, employment, and unemployment. The underground economy, theft losses, and statistical discrepancies have not yet been integrated with criminology theory, for example, but many possibilities exist for going much further in the endogenous treatment of important variables. Much more of the public sector is now endogenous than in the earliest Keynesian models. The social science, legal, and engineering aspects of models need fuller integration and are likely to be handled that way in the future.

REFERENCES

Andersen, L.C., and Carlson, K.M.

1976 St. Louis model revisited. Pp. 46–69 in L.R. Klein and E. Burmeister, eds., *Econometric Model Performance*. Philadelphia: University of Pennsylvania Press.

Ball, R.J., ed.
1973 *International Linkage of National Economic Models*. Amsterdam: North Holland.

Barten, A.P.
1981 COMET in a nutshell. Pp. 211–219 in R. Courbis, ed., *Commerce International et Modèles Multinationaux*. Paris: Economica.

Bodkin, R.G., Klein, L.R., and Marwah, K.
1980 Macroeconometric Modelling: A Schematic History and a View of Its Possible Future. Paper presented to the World Congress of the Econometric Society, Aix-en-Provence, France.

Clark, C.
1949 A system of equations explaining the United States trade cycle, 1921 to 1941. *Econometrica* 17 (April):93–124.

Dramais, A.
1981 Le modèle DESMOS. Pp. 221–234 in R. Courbis, ed., *Commerce International et Modèles Multinationaux*. Paris: Economica.

Duesenberry, J., Eckstein, O., and Fromm, G.
1960 A simulation of the United States economy in recession. *Econometrica* 28 (October):749–809.

Fair, R.
1976 An evaluation of a short-run forecasting model. Pp. 27–45 in L.R. Klein and E. Burmeister, eds., *Econometric Model Performance*. Philadelphia: University of Pennsylvania Press.

Greenberger, M., Crenson, M., and Crissey, B.L.
1976 *Models in the Policy Process*. New York: Russell Sage.

Haavelmo, T.
1943 The statistical implications of a system of simultaneous equations. *Econometrica* 11 (January):1–12.

Hicks, J.R.
1937 Mr. Keynes and the "classics": a suggested interpretation. *Econometrica* 5 (April):147–159.

Klein, L.R., and Goldberger, A.S.
1955 *An Econometric Model of the United States, 1929–1952*. Amsterdam: North Holland.

Klein, L.R., Ball, R.J., Hazlewood, A., and Vandome, P.
1961 *An Econometric Model of the United Kingdom*. Oxford: Basil Blackwell.

Klein, L.R., Pauly, P., and Voisin, P.
1982 The world economy—a global model. *Perspectives in Computing* 2 (May):4–17.

Lange, O.
1938 The rate of interest and the optimum propensity to consume. *Econometrica* 5 (February):12–32.

Mann, H.B., and Wald, A.
1943 On the statistical treatment of linear stochastic difference equations. *Econometrica* 11 (July–October):173–220.

McNees, S.K.
1973 A comparison of the GNP forecasting accuracy of the Fair and St. Louis econometric models. *New England Economic Review* (September–October):29–34.

Modigliani, F.
1944 Liquidity preference and the theory of interest and money. *Econometrica* 12 (January):45–88.

Orcutt, G.H., Greenberger, M., Korbel, J., and Rivlin, A.M.
1961 *Microanalysis of Socioeconomic Systems: A Simulation Study*. New York: Harper.

Polak, J.J.
1953 *An International Economic System*. Chicago: University of Chicago Press.

Suits, D.B.
1962 Forecasting with an econometric model. *American Economic Review* 52 (March):104–132.

Tinbergen, J.
1939 *Statistical Testing of Business-Cycle Theories, II, Business Cycles in the United States of America*. Geneva: League of Nations.

Numbers and Decisionmaking

Public Statistics and Democratic Politics

KENNETH PREWITT

If, to paraphrase Harold Lasswell, politics has become how much for how many, it is clear that measurement moves toward the center of political life. The result is a politics of numbers—What is to be counted? By whom? Can the numbers be trusted? In which direction is the trend line moving? Who is at fault for the (now numerically defined) failure of a policy or program? The intrusion of numbers into politics is global, as the world's nations now endlessly debate issues couched in numerical estimates and forecasts: weapon counts, oil reserves, trade balances, North-South inequities, debt ratios.

With reason, then, scholars have focused their attention on how numbers are generated and subsequently used or misused in politics. This important scholarship rests on the assumption that public statistics are not politically neutral. Decisions about what to count are influenced by the dominant political ideologies, and numbers enter the political fray on behalf of social interests.

The approach adopted in this essay accepts this assumption but focuses it as follows: public statistics in the United States are generated as part of democratic politics. This invites inquiry into the ways in which this particular nation's ''number system'' advances or retards democracy, informs or distorts civic discourse, helps or hinders political participation. For just as public statistics are not neutral with respect to the everyday politics of group interests, so they are not neutral with respect to the principles and practices of democracy. Consequently, to study constitutional democracy, as it is today practiced in the United States, requires a perspective on numerical reasoning and the nation's number system. Providing this perspective is a task for social theory.

There are of course unresolved issues in what does, or should, constitute democracy in the United States. We cannot attempt here to sort out the relative emphasis that contending theories of democracy give to such issues as popular participation, economic and social equalities, the protection of property, civil liberties and citizen rights, or democratic procedures. In this chapter we take the simpler route of concentrating on two central issues: accountability—how public leaders are held accountable for their performance in office; and representation—how diverse interests are represented in setting the political agenda.

DEMOCRATIC ACCOUNTABILITY

The centrality of the concept of accountability in democratic theory derives from the observation that democracies no less than other forms of government have public officials with immensely more power than average citizens. Democratic theory does not deny the power advantage enjoyed by those in charge of the government, nor does it optimistically presume that democracies are free of the tendency of power-holders to expand their control. Embedded in a democracy, again no less than in other forms of government, is a structure of bureaucratic and political power.

The task of democratic theory is to direct us toward practices that reconcile the inclination of power systems toward dominance with the democratic ideal of popular sovereignty. The basic terms of this reconciliation are to be found in the Constitution, especially in the provision for separating and fragmenting official power so that leaders can check and control each other, and in the companion provision that regular electoral competition will force leaders to contest with each other for the favor of the voters.

The general idea of this second provision is summarized in the phrase "theory of electoral accountability" as first adumbrated in *The Federalist Papers* and subsequently elaborated by Schumpeter and other democratic theorists. There is competition for public office. Leaders present themselves and their records to the electorate. Voters, basing their judgments on the past performance or estimates of future performance of leaders, elect, reelect, or evict accordingly. Leaders, knowing this, and wanting to gain and retain office, promote policies that will attract public support.

This theoretical formulation is a reasonably accurate though partial description of what, in fact, does happen. The empirical evidence has been most compellingly presented by political scientist Morris Fiorina (1981; also Kramer, 1971), who has demonstrated the use voters make of retrospective evaluations. Voters routinely reject incumbents who governed during a period marked by deterioration in social and economic conditions. Another political scientist (Kiewiet, 1983:115) reports that voters "clearly react in an incumbency-oriented fashion to the record of current officeholders, responding positively to success in the economic and other arenas but negatively to perceived failures."

Two explanations are available. Citizens vote according to recent changes in their own economic conditions, or, citizens vote according to the improvement or deterioration of national economic conditions.

Under the first explanation, votes would be influenced by the personal experience of unemployment, the personal loss of purchasing power through inflation, or the need to postpone homeowning because of high interest rates. The citizen experiencing these negative economic conditions votes against the political party in power.

Under the second explanation, voters punish or reward politicians depending on the performance of the national economy during the incumbents' tenure. For example, citizens who may be secure in their own employment nevertheless vote against leaders whose policies bring about high rates of unemployment. Or, citizens not themselves seriously affected by high interest rates nevertheless take into account double-digit interest rates when evaluating the performance of incumbent officials.

Somewhat counterintuitively, current research supports the second explanation. Voters in the United States give more weight to negative or positive trends in national economic conditions than to changes in their own economic circumstances. The most extensive development of this finding is offered by Roderick Kiewiet, who concludes (1983:131): "Changing perceptions of the national economy account for a considerably larger proportion of the swing in support for the incumbent party from good years to bad than do changes in personal economic conditions."

This research finding is important in the present context for what it indicates about the function of national statistics in implementing democratic accountability. If voters punish and reward officeholders less in terms of personal experience than in terms of national economic performance, they can vote responsibly only if they have some reasonably accurate information about that performance. This information of course is often made accessible when it is summarized as statistical trends. Political leaders can be judged by the upward or downward movement of statistical indicators of those socially important issues for which government has assumed responsibility: unemployment, inflation, balance of trade, interest rates, test scores, poverty levels, crime rates. When economic and social indicators are moving in politically popular directions, political credit is claimed; when they are moving in unpopular directions, political blame is assigned. Here, then, is a contribution of public statistics to the workings of democracy.

This application of numbers to the purposes of democratic accountability occurs in a period when many other political developments undermine the conditions necessary for holding public officials to account: the decline of party discipline, even of political parties themselves; the increased costs of electioneering and the related packaging of candidates by media experts; the growing political influence of single-issue interest organizations; the comparatively low rates of political participation. These trends occur as the political agenda is ever more crowded with issues difficult for the average citizen to comprehend. A weakened party-electoral system combined with a crowded and complicated issue agenda is not conducive to democratic accountability. Against this background, it is all the more important to understand whether numeric descriptions of major social conditions and trends can improve the reasoning capacity of modern democracies.

The hypothesis can be generalized. Just as a particular administration in power can be evaluated by statistical trends, so also can broad social policies. In this generalized version, citizens continually evaluate and reevaluate broad policy commitments made by previous political generations. In modern nation-states, this retrospective public reflection is facilitated by measures of long-term trends. Descriptive statistics, especially when presented as trend lines, offer voters before-and-after information about the performance of incumbents as well as of general policies. Consequently, these statistics contribute to the procedures that establish accountability in democratic politics. If we could leave matters at this point, the story would be a welcome one for democratic theory. But it is more complicated.

Numbers, just as much as words, have the power to distort as well as enhance the reasoning capacity of the public. The greater the importance of numbers to the securing of power, the stronger the incentives to those in power to make certain that numbers present a favorable even if inaccurate picture. Across a broad front democratic politics must contend with ways in which numbers distort and mislead.

At this point it is necessary to draw attention to an important subset of any nation's number system, what are called "performance indicators." Performance indicators typically serve two functions: they act as internal signals for the agency, telling it whether its goals are being achieved; and they serve as signals to those outside the agency, including of course those who set policy and control budgets. These two functions subject an agency to conflicting pressures. When an agency designs performance measures in a manner that maximizes internal information, it invites external attention to its failures as well as achievements. It risks sending negative signals that those having power over the agency can use to trim budgets or punish incompetence.

It is a familiar complaint that when officials are rewarded or punished according to statistical evaluations, they are drawn to policies that favor how the agency presents itself to the oversight process rather than policies that improve the conditions for which they have responsibility. The numbers become more important than the progress toward policy goals they presumably index. Khrushchev (1977:71) is said to have lamented: "It has become the tradition to produce not beautiful chandeliers to adorn homes, but the heaviest chandeliers possible. This is because the heavier the chandeliers produced, the more a factory gets since its output is calculated in tons."

Our interest is not in this well-known flaw in command economics, but in the implications for democratic accountability. If the number system is systematically manipulated so that personnel and policies are presented to the public in the most favorable light, we have little warrant for claiming that public statistics enhance democratic procedures.

We come here to a point in the discussion where the larger analysis of democratic accountability intersects with a more specific argument about the professional accountability of those who administer the nation's statistical system. This accountability is to professional peers who evaluate, against the standards of their disciplines, whether government statistical agencies are maintaining the integrity of the numbers. Professional statisticians, in and out of government, hold that proper controls and procedures can protect the public from the abuses associated with fraudulent or misleading statistics. In recent congressional testimony, Courtenay Slater, an informed and experienced observer of national statistics, comments (1983:54): "One of the finest things about our statistical system is that our statistics have credibility. They are produced by professionals in the statistical agencies and the press release that gives us our economic data comes out of that statistical agency. It is written by professionals. Everybody knows it is objective, and they believe the numbers are honestly presented."

There is no question that in well-established statistical agencies, such as the Bureau of Labor Statistics and the Bureau of the Census, the production and reporting of statistics is managed by professionals. The norms of professional control are deeply rooted in the origins of these agencies. Consider, for example, the history of the Bureau of Labor Statistics, which just celebrated its centennial. The bureau was established during the post–Civil War period of intense civil strife, and the statistics it produced were quickly implicated in the arguments of the day about the causes and consequences of industrial conflict. Labor reformers especially felt that if information about prevailing employment conditions could be put "before the legislators and the public, a cry of mingled surprise, shame, and indignation will arise that will demand an entire change in the method of earnings and pay" (Wright, 1870:38).

Although not disputing this assumption, the statisticians and socia entists, who by now had organized themselves as the American S Science Association, wanted to ensure that the new labor statistics r seen to favor any particular economic interest—radical, reformist, or servative. The advice given to and heeded by Carroll Wright, found 1873 of the first state Bureau of Labor Statistics in Massachusett subsequently first commissioner of the Federal Bureau of Labor Stati is instructive (Walker, 1877:vii-viii).

> Your office has only to prove itself superior alike to partisan dictation and seductions of theory, in order to command the cordial support of the press and th of citizens. . . . I have strong hopes that you will distinctly and decisively disc the [bureau] from politics.

This advice is no less heeded today than it was a century ago, ar the same reasons. Perhaps no stronger testimony to the credibility o major statistical series is needed than the reliance placed on them by the political process and the marketplace. A member of Congress (H 1983:2) comments, "It would be a public administration catastrophe were to find that the statistics we rely on so heavily did not adequ describe the real world of which we are a part and the problems w trying to solve." In the marketplace substantial funds are routinely ferred on the assumption that national statistical series are trustworthy monthly statistical reports of the Crop Reporting Board of the Depar of Agriculture, for instance, have such high credibility that hundre thousands of dollars change hands through the commodity markets as as the data are released.

But even if we accept that professional control over national sta can largely eliminate fraud and greatly lessen bias in the most imp of our social and economic indicators, other issues remain. The sta of even the most professional agencies suffer from measurement pro for which there are no presently available solutions. When these pro lead to errors of serious magnitude and yet the numbers are used by po leaders to set policies and by citizens to evaluate these policies, th countability process is compromised.

The aptly labeled "unobserved economy" offers a telling illustr Two scholars report (Alford and Feige, n.d.:14): "Recent research su that systematic biases associated with a large and growing sector o measured economic activity have been introduced into the system of indicators. The unobserved sector escapes the social measurement app because of accounting convention, nonreporting and underreporting the unobserved economy is growing more rapidly than the observed is, "counted" economy, but policy is guided by statistics only abo latter, serious errors can hardly be avoided. This in turn of course distorts the process by which fault is assigned, and moves us away from democratic accountability.

The technical and conceptual errors associated with measurement are serious, but for important economic and social indicators continuous professional attention and public discussion offer safeguards. The scholarly community here carries a major responsibility. Social scientists and professional statisticians have the technical skill—and career incentives—to discern discrepancies between what the statistics purport to measure and what they actually measure.

These safeguards can operate only when the statistics are indeed public, that is, accessible to professional attention. Such is not the case for critical domains of national security policy, where secrecy prevails. Professional review of the adequacy and integrity of, say, unemployment or inflation measures is orders of magnitude more informed than professional review of numbers purporting to describe, for instance, the comparative weapon systems of the United States and the Soviet Union.

In a telling essay, McGeorge Bundy (1984) compares statistics on U.S. nuclear weapons appearing in two publications: the officially produced Defense Department *Annual Report for the Fiscal Year 1985* and a privately sponsored report of the Natural Resources Defense Council, the *Nuclear Weapons Databook.* The official publication consistently underestimates American resources in a manner, Bundy argues, designed to make the "Russians look big" and the "Americans small." This is the success indicator issue stood on its head, similar to when police departments inflate crime statistics to justify larger budgets.

At issue is not the inevitable tension between the claims of national security and the right of the public to be informed, for we refer here only to those numbers that are presented by the government in public discussion, from the controversial "body counts" in the war of attrition in Vietnam to the equally controversial "missile counts" in the debates about the window of vulnerability. In sharp contrast to the care with which major statistical series affecting social and economic policies are professionally monitored, there has been little serious attention given to how independent professional controls can be applied to military numbers routinely advanced in open forums.

A democratic society is preserved when the public has reliable ways of knowing whether policies are having the announced or promised effect—Is inflation being brought under control? Is a war of attrition being won? Are defense expenditures buying national security? Numbers, a part of this publicly available political intelligence, consequently contribute to the accountability required of a democracy.

Flaws in the statistics, whether inadvertently or deliberately introduced, mislead citizens about the performance of their government, thereby diminishing accountability, but it can be plausibly argued that the wide public availability of reasonably accurate statistics about social conditions for which government is responsible enhances more than it diminishes democratic accountability. This conclusion, at best an informed guess, rests on assumptions about what is required if civic discourse is to be reasonably informed under the conditions of advanced industrial societies. It also rests on (largely untested) assumptions about the capacity of an electorate to make intelligent use of statistical information. This last point we briefly return to in the concluding section, after reviewing the second of our two major issues connecting national statistics with democratic theory.

REPRESENTATION OF DIVERSE INTERESTS

As a document in democratic political theory, the Constitution's genius is in its provision for the representation of diverse interests in political decision circles. This commitment to representation involved the founders in political engineering, one aspect of which established the close association between political representation and the nation's number system. In order that seats in the House of Representatives might be fairly allocated, the Constitution mandated a population count. It further directed that this count distinguish among the free citizens, the slave population, and the untaxed Indian population. This distinction arose because the founders wanted wealth as well as property to be reflected in apportionment—counting slaves as three-fifths of a person was a way to recognize their property value. Representation had to be apportioned according to politically acceptable criteria. Moreover, the method chosen had to allow for adjustments as the population expanded, redistributing itself among the existing states or spilling over into territories that would later achieve statehood. Thus was established the decennial census, the centerpiece of our statistical system.

The limited use of the census to apportion congressional seats did not satisfy James Madison. In early congressional debates Madison (1790:1077) urged that the census "embrace some other objects besides the bare enumeration of the inhabitants." Madison suggested that the census describe "the several classes into which the community is divided." On this basis, continued Madison, "the Legislature might proceed to make a proper provision for the agrarian, commercial, and manufacturing interests, but without it they could never make these provisions in due proportion."

We know from *The Federalist Papers* that Madison viewed society as consisting of multiple and diverse interests. To govern such a society in a democratic fashion required complex information about the composition of the public. Thus, for Madison, it was not enough that the census enumerate the population for the sole purpose of apportioning. It should be expanded to include many population characteristics, and thereby become the basis on which the legislators could allocate taxes, benefits, and services according to the ''real situation of our constituents.'' In anticipating a democracy in which numerical proportionality cuts much deeper than assigning congressional seats, Madison was ahead of his time.

Madison's opponents started from a different theory of politics. Reflecting eighteenth century theories of the organic society, they ''viewed the object of government as the pursuit of an undifferentiated common good; for them, politics was a sphere of virtue, and empirical investigation was irrelevant'' (Starr, 1984:37). In the early days of the republic Madison's opponents prevailed. Enumeration was sufficient to serve representation.

Contemporary practice, however, is much closer to Madisonian pluralism, as reflected in the vast expansion of the national statistical system and the policy uses to which it is put. The question before us now is how these developments in the statistical system affect the political representation process.

Providing for the representation of diverse interests in political decision circles is at the core of the theoretical formulation known as democratic pluralism, the now dominant interpretation of American democracy. Democratic pluralism takes as its central problem the conditions that allow for the participation by interested parties in various policy domains. Democracy requires that there be no barriers to the organization and expression of the full array of interests in society.

Democratic pluralism is an attractive theory. Since the early days of the republic it has gradually gained adherents among those who have puzzled over the prospects for democracy in large-scale advanced industrial nations. But the theory also has its critics. In recent decades the effort to formulate a democratic theory has emphasized participation as opposed to pluralism, and in the process generated a critique of conventional pluralist theory.

This critique holds that pluralism has not offered a satisfactory account of nonparticipation in democratic politics, too readily attributing low levels of participation to presumed citizen defects such as apathy or ignorance. Since levels of participation covary with social and economic resources, the critics argue, pluralism functions as a justification for the representation of middle and upper class interests in politics rather than a description of how the full array of social interests find a political voice.

An alternative explanation of nonparticipation is suggested by E. E. Schattschneider's famous phrase, ''mobilization of bias.'' In explaining why the socially and economically disadvantaged often fail to participate game is about also decides who gets in the game." This introduces the argument that what is on the political agenda provides a referent point that selectively mobilizes participation across different social groups and interests. Citizens participate not just to put issues on the political agenda but also, and more often, in response to the issues already there. This mobilization process, according to Schattschneider, is biased against the interests of the less well off groups in society.

It is in this theoretical context that we consider how the analysis and political reporting of social statistics intersects the representation system. Although our emphasis is on contemporary politics, the practice we draw attention to is at least 150 years old. Starting around 1820, writes the historian Patricia Cohen (1982:169), "Many private agencies and volunteer groups with reformist agendas adopted the statistical approach to social facts in order to document the dimensions of the problem they were dedicated to eradicating." Cohen offers several examples: the use of statistics to describe the miseries of public prisons; the effort by the temperance movement to prove quantitatively that alcohol abuse was a growing problem; and local surveys of pauperism as a basis to challenge poor laws.

In deploying their privately collected statistics on behalf of social reform, the early nineteenth century activists anticipated developments surrounding publicly collected statistics that did not come fully into view for another half-century, when the federal Bureau of Labor Statistics was established in the 1880s. The 1820 reformers were signaling to later activists that statistics could mobilize political participation and inform public debate.

In the latter part of the twentieth century these possibilities are etched much more deeply in our political life. The nation's number system uncovers social conditions and popularizes them as statistical descriptions: proportion of the population below the poverty line; incidence of child abuse; persistence of structural unemployment; addictive behavior and its social costs; the differential in infant mortality between whites and nonwhites; the gap between male and female wages in similar occupations. The transformation of politically unnoticed social conditions into visible statistics puts issues on the political agenda that would otherwise be ignored.

These statistical conditions then provide a political referent point for interested groups. This is perhaps one of the most striking aspects of twentieth century democratic politics. Resource-poor social interests turn to a statistical description of their plight to generate political pressure and to mobilize adherents to their cause.

The history of the civil rights movement is suggestive in this regard. The concept of institutional racism, which held that black poverty was caused not just by racial prejudice but also by structural conditions of the economy, polity, and society, made its political appearance through statistics on residential segregation, black-white income differentials, unequal educational opportunities, inequities in access to health care, and so forth. Civil rights leaders first used the numbers to emphasize the scope of institutional discrimination. They then used them to gain political support for new social policies such as Headstart, job training, and affirmative action. Other groups have reached the conclusion that to be ''measured'' is to be politically noticed, and to be noticed is to have a claim on the nation's resources. Thus the physically handicapped in New York initially resisted being counted, for fear that this would lead to further stigmatizing them, but then reversed their position when they realized that political visibility closely followed on statistical visibility.

Data presented in Michael Harrington's *The Other America* helped initiate the War on Poverty by identifying the poor as a target group for government action. The consumer protection movement, starting with Ralph Nader's *Unsafe at Any Speed*, has made heavy use of statistical arguments, as have the environmentalists. Describing public-interest citizen groups, one commentator (Henderson, 1981:441) writes that ''the quality and quantity of information and the way it is structured, presented and amplified'' shapes their political choices and strategies.

Harold Wilensky (1967:19) generalizes these observations when he writes that ''facts and figures'' assist those political interest organizations ''weak in grass-roots political resources.'' Information ''may give an advantage to the weak, whose case, if strong and technical, can count for something.'' This is not a trivial observation when examined in the context of the effort through the history of democracy to establish equal civil and political rights in the face of inequalities in resources that different social interests bring to the political arena.

In democratic theory as well as actual practice, organization is most often promoted as the corrective when economic inequalities are reproduced as differential opportunities for political participation. The less wealthy but more numerous social interests combine and increase their political strength through working-class parties, social movements, and interest groups. Consequently, a resource that helps to organize the resource-poor will help to correct political imbalances and promote broader democratic participation.

This observation leads us to consider whether statistical programs can actually help establish group identity and lead to the formation of interest organizations. In a careful account of the interplay between ethnicity and the census, William Petersen (1983:27) writes: ''Few things facilitate a category's coalescence into a group so readily as its designation by an official body,'' and cites the importance of ''questions put to them by immigration officials and census schedules'' for helping to solidify group identification.

Hispanic-Americans are particularly important in this regard. More than any group in American political history, Hispanic-Americans have turned to the national statistical system as an instrument for advancing their political and economic interests, by making visible the magnitude of the social and economic problems they face.

In the processes by which groups are formed and diverse interests are represented in democratic politics, public statistics are not an unmixed blessing. Just as some groups can establish a political identity by being enumerated, other groups cannot escape the way they are socially classified because of this same enumeration system. For example, for two centuries we have had a statistical practice of racial classification, which undoubtedly has contributed to the continuing salience of race in American society. Policies now being implemented could easily result in the Hispanic-Americans becoming a permanent racial minority in the statistical system, with what long-term effects it is difficult to foresee. Moreover, the statistical system is not sufficiently robust to withstand the distortions accompanying severe political pressures. When political criteria are transparently used to determine what should be technical issues, such as the best way to count a population group, statistics lose their credibility.

Racially sensitive measurement policies are not likely to be reversed soon, now that so many government services are allocated according to race and ethnicity. The brief period during which it was thought wrong to identify race, gender, or national origins on employment or school applications was swept away by the emergence of affirmative action and statistical parity in the 1970s. The nation has entered a period in which ''proportionate allocation'' is carried to ever greater extremes. There is a contagion effect: Once statistical proportionality is elevated to a principle of government, there is great pressure from various racial and ethnic groups to be fully counted.

From the perspective of democratic theory these developments are troubling in at least three respects. First, to assign to the statistical system responsibility for group classification and resource allocation is to transform the thing being measured—segregation, hunger, poverty—into its statistical indicator. Always in tension with the judgmental in politics is an insistent search for objective rules to reduce the element of arbitrariness in subjective judgment. The legal code is one such set of objective rules, formalized bureaucratic procedures another, and now statistical formulas. This search does not eliminate politics; it simply pushes them back one step, to disputes about methods. Arguments about numerical quotas, availability pools, and demographic imbalance become a substitute for democratic discussion of the principles of equity and justice.

Second, if statistical identification facilitates political consciousness among some resource-poor groups, these same statistics make invisible to the policy process other groups at the margins of social and economic life, where measurement often fails—the undocumented workers, the illegal aliens, and the vagrant, homeless populations. In many government programs, persons not counted are not there. Another difficulty stems from the inertia of statistical systems. For technical as well as bureaucratic reasons, statistics lag behind the dynamic patterns of group formation and change resulting from immigration, internal migration, transformation in the occupational structure, and new levels of social consciousness. Insofar as politics is organized by the numbers, there will be a tendency to overlook more recently established social conditions in favor of those already reflected through the statistical system.

The third and most troubling danger is the shift away from a system of representation and public policy based on the individual citizen toward one based on the representation of demographic aggregates: ethnic, racial, income, gender, etc. This shift invites, even mandates, the allocation of benefits and rights according to group membership rather than individual accomplishment or need.

To many observers this tilt toward group representation undermines the fundamental premise of liberal democracy. Nathan Glazer (1975:220) laments the drift toward numbering and dividing up the population into racial and ethnic groups: "This has meant that we abandon the first principle of liberal society, that the individual and individual's interests and good and welfare are the test of a good society, for we now attach benefits and penalties to individuals simply on the basis of their race, color, and national origin." Glazer, of course, does not attribute the rise of quota politics and group-based representation to the availability of statistical information. But if statistical information has not caused, certainly it has abetted the emergence of demographically defined groups as a category in public policy.

The formal system of political representation itself has not escaped the insistent pressure for demographically defined proportionality. Abigail Thernstrom (1983) artfully traces how the 1965 Voting Rights Act was transformed in two decades from a law to protect black voting rights to one that appears to require the "correct" number of minority seats in legislative bodies. Demands for proportional representation, in which the legislature is to mirror the characteristics of the population from which it is selected, are not new. Until recently, however, group politics intersecting with the electoral process was the preferred avenue for achieving this end. Legal remedies were, appropriately, limited to ensuring fair procedures, not particular outcomes. Now, buttressed by statistics, laws have begun to affect the very composition of legislative bodies.

As was the case in our discussion of accountability, we see in this discussion of representation that countertendencies are at work. On the one hand, statistical description can bring social conditions to public attention, mobilize disadvantaged groups, and broaden the political agenda in ways that lessen the bias inherent in an electoral representation system based largely on the resources of wealth and political organization. On the other hand, these statistics introduce practices and policies inconsistent with our traditional understanding of democracy: the objectification of politics; the assumption that that which is not counted is not there; the temptation to substitute group membership for individual merit or need as the basis for public policy; the allocation of legislative seats according to designated racial or ethnic criteria.

We are far from having the evidence that would allow us to sort out the relative strength of these countertendencies and again must resort to an informed guess. With respect to democratic accountability I suggested that the benefits of statistical descriptions outweighed the harms. With respect to the representation of diverse interests I am less sanguine. The distortions of the representational process seem to me every bit as strong as the improvements. Moreover, the negative tendencies are not of the sort that can be corrected with greater professional scrutiny of statistical information. They are much more political than technical in nature and in fact become stronger as statistics become more precise and reliable.

CONCLUSIONS

I conclude by emphasizing the theme that connects this essay with the efforts of Ogburn and colleagues. The present inquiry has emphasized the importance of close attention to the nation's number system by professional statisticians and social scientists. Assuring the integrity of numbers involves continuous improvements in measurement, revisions in concepts as social conditions change, and the highest standards of statistical interpretation, analysis, and reporting. Moreover, protecting statistical quality and integrity will add little to democracy unless joined to the educational task of ensuring that numeracy takes its place alongside literacy as a skill indispensable to democratic citizenship. In the absence of public understanding of statistical argumentation, the numbers will more likely aid political demagoguery than democratic discourse.

Because of, and notwithstanding, the various problems and risks identified in this essay, those who care about democracy have a large task before them: analysis of the political role of numbers, as well as a commitment to making the numbers perform according to the responsibilities that a democracy places upon them.

ACKNOWLEDGMENTS

To fully acknowledge collegial contributions to this chapter would require this note to be nearly as long as the paper itself. It was prepared while the author was a Fellow at the Center for Advanced Study in the Behavioral Sciences. The working conditions at the center are much appreciated; colleagues at the center who provided useful commentary include Joel Aberbach, Paul Brest, Roger Noll, and Barbara Rosenkrantz. I am grateful for the support of the Guggenheim Foundation, and also for the financial support provided through the center by the National Science Foundation, under grant BNS 76-22943, and by the Exxon Educational Foundation. I appreciate as well critical comment from participants at the symposium on which this volume is based and especially from anonymous commentators commissioned to review the very preliminary draft presented at the symposium. In addition, this chapter benefited from the Social Science Research Council's Conference on the Political Economy of National Statistics. A slightly modified version is included in a volume based on that conference, which gave me access as well to the editorial advice and sharp editorial pens of William Alonso and Paul Starr.

REFERENCES

Alford, Robert R., and Feige, Edgar L.
N.d. Information Distortions in Social Systems: The Unobserved Economy and Other Observer-Subject-Policy Feedbacks. Unpublished paper.

Bundy, McGeorge
1984 Deception, self-deception and nuclear arms. *The New York Times Book Review*, March 11:3,15.

Cohen, Patricia Cline
1982 *A Calculating People: The Spread of Numeracy in Early America*. Chicago: University of Chicago Press.

Fiorina, Morris
1981 *Retrospective Voting in American National Elections*. New Haven: Yale University Press.

Glazer, Nathan
1975 *Affirmative Discrimination: Ethnic Inequality and Public Policy*. New York: Basic Books.

Henderson, Hazel
1981 Information and the new movements for citizen participation. Pp. 434–48 in Thomas J. Kuehn and Alan L. Porter, eds., *Science, Technology, and National Policy*. Ithaca, N.Y.: Cornell University Press.

Horton, Frank
1983 *Federal Government Statistics and Statistical Policy*. Hearing before the Legislation

and National Security Subcommittee of the Committee on Government Operations, House of Representatives, June 3. Washington, D.C.: U.S. Government Printing Office.

Khrushchev, N.

1977 Cited in Charles E. Lindblom, *Politics and Markets*. New York: Basic Books.

Kiewiet, D. Roderick

1983 *Macroeconomics & Micropolitics: The Electoral Effects of Economic Issues*. Chicago: University of Chicago Press.

Kramer, Gerald

1971 Short-term fluctuations in U.S. voting behavior, 1896–1964. *American Political Science Review* 65:31–43.

Madison, James

1790 *Annals of Congress*, First Session, House of Representatives. Washington: Gales and Seaton, 1834. Cited in Steven Kelman, The politics of statistical policymaking: justification for public information-collection and theories of the role of government, Conference on the Political Economy of National Statistics, October 14–15, 1983:14. Social Science Research Council, New York. Revised version to be published in William Alonso and Paul Starr, eds., *Politics of Numbers*.

Petersen, William

1983 Politics and the measurement of ethnicity. Conference on the Political Economy of National Statistics, October 14–15. Social Science Research Council, New York. Revised version to be published in William Alonso and Paul Starr, eds., *Politics of Numbers*.

Schattschneider, E.E.

1960 *The Semi-Sovereign People: A Realist's View of Democracy in America*. New York: Holt, Rinehart and Winston.

Slater, Courtenay

1983 *Federal Government Statistics and Statistical Policy*. Hearing before the Legislation and National Security Subcommittee of the Committee on Government Operations, House of Representatives, June 3. Washington, D.C.: U.S. Government Printing Office.

Starr, Paul

1984 Measure for measure. *The New Republic* February 13.

Thernstrom, Abigail

1983 Trigger of reform, trigger of mischief: the use of statistics in voting rights and redistricting. Conference on the Political Economy of National Statistics, October 14–15. Social Science Research Council, New York. Revised version to be published in William Alonso and Paul Starr, eds., *Politics of Numbers*.

Walker, Francis A.

1874 *Massachusetts Bureau of Statistics of Labor, Fifth Annual Report*. Cited in James Leiby, *Carroll Wright and Labor Reform: The Origin of Labor Statistics*. Cambridge: Harvard University Press, 1960:63.

Wilensky, Harold

1967 *Organizational Intelligence*. New York: Basic Books.

Wright, Carroll

1870 *Massachusetts Bureau of Statistics of Labor, First Annual Report*. Cited in James Leiby, *Carroll Wright and Labor Reform: The Origin of Labor Statistics*. Cambridge: Harvard University Press, 1960:55.

1877 *Massachusetts Bureau of Statistics of Labor, Eighth Annual Report*. Cited in James Leiby, *Carroll Wright and Labor Reform: The Origin of Labor Statistics*. Cambridge: Harvard University Press, 1960:68.

Deterrence in Criminology and Social Policy

H. LAURENCE ROSS and GARY D. LAFREE

INTRODUCTION

Social policy and research on the deterrence of crime have often been unrelated in the United States. While politicians have periodically called for harsher punishments to deter crime, most criminologists prior to the 1970s either ignored the deterrence issue or voiced strong skepticism toward it (e.g., Sutherland, 1924:360; Reckless, 1967:504). This gap between policy and research is unfortunate, manifesting itself in policy initiatives unrefined by empirical evaluation and empirical research with little policy significance.

In recent years the estrangement between criminology and social policy on deterrence has shown signs of abating. This chapter examines recent social research on two important categories of modern human misconduct—street crime and drunk driving—to explore the implications of recent criminological studies for policy on the deterrence of crime.

TWO FUNDAMENTAL PERSPECTIVES ON CRIMINAL CONDUCT

Two broad perspectives on human behavior have long competed for preeminence in efforts to control crime in America. Both have historical roots as well as present-day champions, and both have been evident in the operation of our legal system since its inception. The first asserts that human behavior may be usefully represented as the product of rational individual calculation; the second asserts that behavior is largely guided by nonrational biological, psychological, or social forces.

The ''rational actor'' perspective assumes that human beings behave to maximize personal pleasure and minimize pain. As elaborated by social reformers like Bentham and Beccaria, or jurists like Blackstone, Romilly, or Feuerbach, this view argues that crime can be deterred by increasing the costs of criminal behavior or increasing the rewards of noncriminal behavior. Contemporary criminologists generally refer to this perspective as the ''Classical School'' (Jeffery, 1972; Vold, 1979).

By contrast, the second perspective assumes that human behavior, including crime, is governed by forces over which the individual has relatively little conscious, rational awareness or control. Starting with Cesare Lombroso and his students in the 1800s, criminologists have labored to discover, describe, and understand these forces. Contemporary criminologists refer to this perspective as the ''Positive School.''

For those who assume that crime is caused by factors outside the offender's control, the proper role of criminology is not to investigate the deterrent effects of variations in the law and its enforcement but, rather, to help ameliorate the problem of crime by identifying and taking steps to alter the biological, psychological, or social conditions that produce it. Whereas the classical perspective suggests the possibility of deterring crime through manipulating actual or expected rewards and punishments, the positive school recommends changing nonrational elements of the offender's psyche or environment. The most common policy approaches for bringing about these changes include a variety of intervention strategies that regardless of their actual performance, have generally been justified as ''rehabilitative'' (e.g., probation, parole, indeterminate sentencing, and institutional treatment).

Public policy on crime in the twentieth century has drawn from both the positive and classical schools. The belief that credible threats of punishment deter criminal behavior is probably as old as criminal law itself and has broad appeal to policymakers and the public. From an intuitive point of view it seems reasonable. Surely Chinese citizens were less likely to exceed the speed limit in Peking early in this century when authorities exhibited the heads of drivers executed for speeding alongside speed limit signs (Zimring and Hawkins, 1973:11). The adoption of harsh laws against crime in the United States, as well as the mobilization of criminal penalties to deal with specific behavioral problems—such as drunkenness and other drug abuse—show continued faith in the efficacy of deterrence. At the same time, twentieth century policymakers have created vast programs to rehabilitate criminals, including probation, parole, and specialized correctional facilities for juveniles and for specific categories of convicted offenders. While some of these efforts at rehabilitation have been called half-hearted, no one can seriously deny that substantial

In contrast to the dual approach of policymakers, balancing (perhaps vacillating) between deterrence and rehabilitation, criminologists have by and large rejected deterrence principles. For over a century a host of distinguished scholars with viewpoints as different as Enrico Fermi and Edwin Sutherland were able to agree on one point: deterrence does not work.[1] One major reason for this long-term rejection of deterrence is that criminologists, especially in the United States, traditionally were humanists and reformers (Gibbons, 1979; Wilson, 1983a). Since the early years of this century American criminology has had a strong social reform component, based on the belief that government is not merely a device to facilitate the pursuit by individuals of their private ends, but also a device to shape and improve the character of its citizenry. The reformers held that if only the right institutions were built, the right people properly trained to staff them, and the right classification procedures used to fill them, then surely rehabilitation would occur.

However, in the mid-1970s a profound retreat from these assumptions became evident among criminologists and policymakers. For the first time in 150 years American criminologists seriously questioned whether rehabilitation was a reasonable goal. State governments across the United States were moving away from indeterminate sentencing (long associated with the rehabilitative ideal), curtailing the use of probation and parole, and speaking against "correctional" programs slanted toward rehabilitation.

The reasons for the recent decline in the popularity of rehabilitation, both among policymakers and criminologists, undoubtedly warrant a separate, detailed account. Here we can merely summarize the more common explanations for the shift. First, renewed interest in deterrence appears to reflect the perceived failure of rehabilitation policies. This failure was typically "proven" by the precipitous increase in crime rates in the last two decades (Wilson, 1975, 1983a, 1983b); by widely publicized prison disasters, such as those in Attica in 1969, and Santa Fe in 1980; and by mounting research evidence that many programs aimed at preventing recidivism through rehabilitation programs have been relatively ineffective (Martinson, 1974; Lipton et al., 1975).

Second, traditional methods of rehabilitation have come under increasingly strong attack in the last two decades on the basis of their intrusiveness. In a series of decisions in the 1960s, most notably *Escobedo v. Illinois* (378 U.S. 478, 1964) and *in re Gault* (387 U.S. 1, 1967), the Supreme Court revolutionized the meaning of due process rights in American law

[1] In fact, this belief was fully articulated by Edwin Sutherland in the chapter he wrote on "Crime and Punishment" for *Recent Social Trends in the United States*, the Ogburn Report.

with important consequences for the idea of rehabilitation, with its emphasis on individualized intervention to bring about psychological changes in offenders. Disdain among both politicians and criminologists for this type of intervention became increasingly evident in the late 1960s. For example, *Struggle for Justice* (1971), the influential report prepared for the American Friends Service Committee, asserted (p. 85) that rehabilitation rests "largely on speculation or on assumptions unrelated to criminality," and that decisions made about offenders are routinely made "in the absence of credible scientific data on the causation or treatment of crime." This critique and others like it attacked the fundamental assumptions of rehabilitation: that crime is caused by forces over which the individual has little control and that the criminal justice system is able to identify and correct these forces.

Finally, a less theoretical but eminently plausible explanation for the decline in support for rehabilitation programs is their cost. Rehabilitation, especially the kind envisioned by much of the criminological literature, is expensive. Many states find themselves spending increasing amounts of scarce revenues on correctional programs that are often difficult to justify to taxpayers. Andrew Scull (1977) and others have argued that these purely economic forces, rather than humanitarian or scientific concerns, have led to declining support for rehabilitation.

With diminishing support for rehabilitation, the justification for punishing criminals has shifted toward deterrence, and the forms of scholarship have likewise changed. It is difficult to find more than a half-dozen professional articles or books written on the subject of deterrence from 1900 to 1965. But starting in the 1960s, the study of deterrence has become a criminological growth industry. Within criminology, the deterrence proposition has generated new interest and a large and rapidly expanding research literature (see Zimring and Hawkins, 1973; Andenaes, 1974; Gibbs, 1975; Cook, 1977; Blumstein et al., 1978; Tittle, 1980; Archer et al., 1983; for reviews). Stated simply, this proposition asserts that *proscribed behavior is deterred by perceptions that legal punishments are swift, sure, and severe.*

Our main purpose in this chapter is to appraise and interpret research on the deterrence proposition. Because researchers know little about the consequences of swift punishment in the context of laws—there is too little swift punishment available in our legal system to study—our exclusive focus is on variations in the certainty and severity of punishment. Moreover, evaluations of the effects of certainty and severity of punishment must distinguish between the legal existence (*de jure*) of punishment, and its actual use (*de facto*). Despite *de jure* changes in punishment, meaningful *de facto* changes have been rare. Thus, evaluations of the deterrence proposition are confounded by the fact that real changes in the imposition of punishment seldom accompany legal changes. We review prior research for two important types of criminal behavior, street crime and drunk driving, present generalizations that have been established in these fields, evaluate the strength of the evidence, and interpret their meaning for social policy.

STREET CRIMES

Perhaps no area of deterrence research has generated as much public interest in recent years as attempts to reduce street crime, which generally means robbery, rape, assault, or murder that occurs in public places between people previously unacquainted. The public is concerned and fearful: the President's Commission on Law Enforcement (1967) found that one-third of all Americans were afraid to walk alone at night in their own neighborhoods, and many reported that they stayed off the streets altogether because of their fear of crime. Subsequent victimization surveys conducted by the Department of Justice (see, e.g., Hindelang, 1976) have confirmed the enormous impact that fear of street crime has on the behavior of citizens—especially the poor, members of minority groups, and urban residents. "Crime in the streets" has been a recurring national and political issue since the mid-1960s, and strategies for dealing with street crime are the subject of debates, media programming, political campaigns, and government commissions. In light of this interest social science knowledge concerning the effect of deterrence-based legal interventions bears important policy implications. This knowledge is summarized here.

Certainty of Punishment

The deterrence proposition predicts that proscribed behavior will be reduced to the extent that the relevant public perceives a high likelihood of punishment for violations.[2] As other reviewers (e.g., Blumstein et al., 1978) have noted, relatively little effort has been made to measure this perception directly; the bulk of what we know simply relates aggregate measures of street crime to policy or program innovations that are *intended* to increase the *actual* chance of punishment, implicitly assuming that increases in the (intended or actual) likelihood of punishment will in turn lead to increases in its *perceived* certainty. This chain of assumptions may reasonably hold where the innovations are accompanied by official publicity and mass media

[2] Deterrence research distinguishes between "general" and "special" types. General deterrence is the inhibiting effect of sanctioning an offender on other potential offenders' criminal behavior. Special deterrence is the inhibiting effect of sanctioning an offender on his or her own future criminal behavior.

attention, but its unproved assertion is a point of weakness in much of the existing evidence.

The best-studied legal innovation to increase the certainty of punishment has been intensive policing efforts designed to raise the risk of apprehension and charging.

Police crackdowns are fairly common responses to public concern about street crimes, and a number of these have been submitted to evaluation. Of particular interest are two studies of patrol efforts in New York City, "Operation 25" and the "20th Precinct" studies; the San Diego Field Interrogation project; a study of robberies in the New York City transit system; the LEAA High-Impact Anti-Crime project; and the Kansas City Preventive Patrol project.

One of the earliest evaluations of increased patrol's effect in reducing street crimes was Operation 25 in New York City (see Zimring and Hawkins, 1973:348–349). The police department selected the twenty-fifth precinct, a small district with a high crime rate, for greatly increased patrol during a four-month experimental period. The number of foot-patrol officers within this district was quadrupled for the experiment, and crime rates declined in all categories during those four months. However, no data were available to investigate the possibility that Operation 25 may have shifted the location of crimes from the experimental precinct to adjacent areas.

This weakness in Operation 25 was avoided in a subsequent, similar study in the twentieth precinct, which received a 40 percent increase in police manpower and also noted decreases in the rates of major crimes (Press, 1971). The evaluation controlled the experimental data with findings from adjacent districts to test for displacement effects and from distant districts to test for the possibility that a general decline in crime rates could have explained the decline observed in the experimental district. The control data supported the conclusion that the patrol was effective in reducing street crimes (i.e., those visible from the street) and that it did not merely displace the criminal activity into adjacent districts. Of course, these results can also be questioned. The period of the experiment was only eight months and only changes in crimes reported to the police were measured. Moreover, the official records were maintained by police who were aware of the experiment and/or the previous findings from Operation 25.

One of the best-designed experiments on the deterrence of street crime, the San Diego project (Boydstun, 1975), concerned the practice of stopping, questioning, and frisking persons who aroused police suspicions (i.e., conducting "field interrogations"). In one area of the city, field interrogations were eliminated, whereupon the number of "suppressible" crimes (robbery, burglary, theft, auto theft, assault, sex crimes, malicious mischief, and disturbances) increased by about a third; when field interrogations were resumed, the number of such crimes dropped back to preexperimental levels. There was no change in the frequency of suppressible crimes in two control areas where either field interrogation practices remained unchanged or police officers were specially trained to conduct them in light of legal procedures and human relations principles. Because the presence or absence of field interrogations did not affect the number of arrests in either control or experimental areas, Boydstun concludes that the visibility of police activity was responsible for the apparent deterrence of crime.

In response to a large increase in subway robberies (especially of toll-booth stations) in 1965, the New York Transit Authority introduced special patrols on the subways during nighttime hours. Evaluators (Chaiken et al., 1974) found that crime rates dropped substantially during the patrol hours, but not during the balance of the day, for up to six years following the crackdown. An interesting sidelight in the study that has important implications for deterrence research was the discovery of a ''phantom effect.'' For eight months, while there were stepped-up patrols only at specific times and places, serious crime rates declined throughout the subway system. The evaluators assert that uncertainty as to the deployment of the police had a deterrent effect on potential offenders in nontarget areas and times. However, in the long run this phantom effect disappeared, possibly because potential felons became familiar with actual deployment practices. One difficulty with this study was that the evaluation was based on crime reports made by the participating transit police officers. After the evaluation was completed, researchers alleged that police officials miscoded the times of some offenses, apparently to exaggerate the reduction in frequency of offenses during peak patrol hours (Gallagher, 1978:176). In a reexamination of the evidence, Chaiken (1976) concludes that despite the falsification, there was a significant deterrent effect, although of lesser magnitude than the original evaluation suggested.

Perhaps the most ambitious experiment in deterring street crime ever attempted in the United States was the High-Impact Anti-Crime Program, funded for $160 million by the Law Enforcement Assistance Administration in 1972 (Chelimsky, 1976). Eight cities with high crime rates were targeted for crime-reduction programs with the goal of reducing stranger-to-stranger personal crime by 5 percent in two years and 20 percent in five years. Each city had complete discretion in designing individual programs and evaluating the consequences. Most of the money was allocated to increased enforcement, although some projects also aimed at streamlining court operations. Unfortunately, the variability in programs and evaluations strongly compromised the scientific utility of the program. A summary evaluation (Chelimsky, 1976) found the individual project descriptions to be flawed that omitted several theoretically crucial variables, including conviction rates, sentencing patterns, and arrest-to-crime rates. As Franklin Zimring points out (1978:144–149), the UCR data do not allow specific measures of stranger-to-stranger offenses, those targeted for reduction by the project. Even so, the final evaluation offered no explicit statistical comparison for crimes other than burglary. The report found that five of the eight Impact cities had 1974 levels of burglary lower than would have been predicted on the basis of extrapolations from a set of comparison cities, but no such differences were apparent in the remaining three cities. Thus, modest support was obtained for the deterrence proposition from this very costly experiment.

The Kansas City Preventive Patrol project was designed to test the relative effect on crime rates of three policing strategies: ''proactive'' patrol, with patrol car levels between two and three times the normal level; ''normal'' patrol; and no routine patrol, police entering the area only in response to calls for assistance (Kelling and Pate, 1974). These strategies were assigned randomly to 15 contiguous districts within the city and were maintained for 12 months. The major dependent variables were official police statistics and victim reports. No deterrent results could be found for patrol at either ''normal'' or ''proactive'' levels.

Critics of the Kansas City study (cited in Zimring, 1978:142–143) have pointed out that the districts were small, and that residents not subjected to patrol could still see police patrolling peripheral areas and responding to calls. There were, in fact, no significant differences in police response time or arrest rates between the three kinds of patrol districts. The comparison between no-patrol and routine patrol areas might have been contaminated by this proximity effect. It is notable that the Kansas City patrols were by car whereas prior studies reporting positive effects used foot patrols. However, the study was widely interpreted as failing to show that doubling or tripling of police patrol could measurably affect crime rates.

Severity of Punishment

The deterrence proposition also predicts that proscribed behavior will be reduced to the extent that the relevant public perceives great severity of punishment for violations. Most research on this hypothesis compares jurisdictions that have death penalty provisions with those having (presumably less severe) prison sentences; compares jurisdictions with longer and shorter prescribed or actual prison sentences for various offenses; or examines longitudinal effects of changes in punishment severity on official crime rates (for reviews, see Andenaes, 1974; Zeisel, 1976; Blumstein et al., 1978). Regardless of methods used, there is little evidence directly bearing generally accord with the severity of formal legal prescriptions. This assumption may be false in specific circumstances.

The Death Penalty Capital punishment obviously cannot be experimentally manipulated, and examinations of its deterrent effect have been limited to comparisons of homicide rates in contiguous states with and without the death penalty (Campion, 1955; Sellin, 1959); to examinations of time-series data on homicide rates within one or more jurisdictions that change capital punishment laws (Sellin, 1959; Walker, 1969); and to comparisons of homicide rates within a jurisdiction before and after the imposition of a death sentence or execution (Graves, 1956; Savitz, 1958). Although these studies have generally failed to find evidence for a deterrent effect of capital punishment, they have serious methodological problems that compromise their probative value. The main problems lie in their inability to control for demographic, cultural, and socioeconomic factors other than the death penalty that could affect rates of serious criminality, and their failure to distinguish between the formal prescription of the death penalty and its actual application.

Recent support for the deterrence proposition in the matter of capital punishment has been reported in a well-known study by Isaac Ehrlich (1975), who examined aggregate U.S. data on homicide rates and capital punishment for the years 1932–1970. After performing an elaborate set of statistical analyses, Ehrlich concluded that capital punishment does deter homicide, offering a specific estimate of the magnitude of the effect (p. 398): "On the average the tradeoff between the execution of an offender and the lives of potential victims it might have saved was of the order of magnitude of 1 for 8 for the period 1933–1967 in the United States."

Ehrlich's study, introduced to the Supreme Court by the Solicitor General in *Fowler v. North Carolina* (95 Sup. Court 223, 1975), has been widely cited in support of capital punishment legislation and its imposition in individual cases. Because of the importance of the findings and the fact that it is one of the few studies to report a deterrent effect for capital punishment, its supporting data have been reanalyzed several times (e.g., Bowers and Pierce, 1975; Klein et al., 1978). These analyses show that Ehrlich's findings are sensitive to minor changes in the form of the analysis. Among the most striking is the consequence of changing the time period over which the analysis is made: the negative relationship between homicide rates and executions is present only when the years 1962–1969 are included in the analysis, and these were unusual years for the United States in that both homicide and all other street crimes increased dramatically while the frequency of executions declined steeply (and ceased entirely in 1968; see

Bowers and Pierce, 1975:197–202).[3] Thus, the probative value of th lich study is doubtful.

The bulk of the existing literature on whether capital punishment crime more than other forms of punishment (e.g., life imprisonm based on data gathered in situations in which few convicted off actually receive capital punishment (and the method for selecting offenders receive it appears highly capricious, despite recent Supreme decisions aimed at clarifying standards), so the generally negative fi must be understood as limited to situations in which actual likelih punishment is low. It is possible that if executions were applied with likelihood they might have an effect on homicide rates, but the wisd such a policy is primarily a moral and ethical matter, not a sci question. In any event, policy decisions about capital punishment ar likely to be affected by moral and ethical considerations than by the estimates of deterrent effects that present social research has produ

Other Punishments The quantity of studies of the deterrent eff noncapital punishments on street crime is somewhat more impressi a review, see Nagin, 1978). However, there is only one study we k where the design rises to the level of a quasi experiment. Schwartz studied the effect of increased statutory penalties for rape and atte rape on the frequency of these crimes in Philadelphia. Following a rape case that received a great deal of media attention, the state of sylvania raised the maximum penalty for rape by a factor of two or depending on the severity of the case. Schwartz examined reporte rates for the period surrounding the change and concluded that neit frequency nor the seriousness of rape changed significantly after th law was passed. The study has obvious weaknesses, most notably t of reported cases of a crime that is notoriously underreported. Howe avoids the general problem that affects the balance of the known s the impossibility of adequately controlling statistically for all the im social, economic, and demographic variables other than deterrence laws that can affect crime rates.

The general picture painted by the bulk of studies relating seve punishment for street crimes as measured by length of sentences (pre or actual) to crime rates is less favorable to the deterrence prop (Chiricos and Waldo, 1970; Forst, 1976). The predominant findin no significant relationship. In an exhaustive review of the evidenc

[3] A detailed methodological critique of Ehrlich's analysis by Klein et al. (1978:343–3

deterrent effect of sentence severity on crime, Nagin (1978:110) concludes that at best the results are "equivocal." As with the death penalty studies, though perhaps to a lesser degree, these studies also take place against the background of a relatively low risk of any punishment. Hence, results more favorable to the deterrence proposition might be found if severe punishments were perceived to be likely in the event of violations by the population to which the threat is addressed.

Summary

On the matter of *certainty* of punishment, we find some support for the deterrence proposition in the literature on street crime. With few exceptions, crime rates are found to decline when measures are adopted to increase the certainty of punishment. However, there is much weaker confirmation for the deterrence proposition in the matter of *severity* of punishment, with a few studies claiming an effect contradicted by numerous studies finding no effect. Because all empirical research in the area of severity of penalties takes place in situations where the objective likelihood of punishment is very low, scientific generalizations and policy decisions based on this literature should be appropriately qualified.

DRUNK DRIVING

Unlike the case for street crimes, active public interest in the issue of drunk driving is relatively recent, and its depth and persistence have not yet been tested. The acute current concern about drunk driving has resulted in a flood of deterrence-based legal interventions that promise to increase our understanding of deterrence both in the specific case and in general. Laws have been passed and enforcement campaigns undertaken with the aim of increasing the perceived certainty and severity of punishments for drunk driving, forming a pool of natural experiments that can be subjected to analysis. Furthermore, the results can be ascertained using indexes such as weekend night fatalities, which are often both validly and reliably measured by official statistics agencies, and which correlate strongly with alcohol-impaired driving. There now exists a relatively large body of knowledge in this area and additional experience is rapidly accumulating.

Certainty of Punishment

Two types of legal interventions regarding drunk driving are directed primarily at increasing the objective (and hence, presumably perceived) certainty of punishment. The first of these is the replacement of laws that define the offense in behavioral terms with "Scandinavian-type" or "per se" laws, which define the offense in terms of blood-alcohol concentrations, measurable by instruments. Typically these laws require only that a driver exceed the tolerated level in order to justify arrest and to secure conviction; there is no need to demonstrate drunken behavior or to produce other evidence of impairment, a matter that under previous law was a severe handicap in detecting and prosecuting drivers whose chances of experiencing a crash were substantially increased by the consumption of alcohol. The second type of intervention is the enforcement crackdown, in which police resources devoted to drunk-driving patrols are abruptly increased. We will review the literature accumulated to date on the deterrent effects of these interventions.

Scandinavian-type Laws These laws originated in Norway in 1936 and Sweden in 1941, where they formed part of general accumulations of legal restrictions on drunk driving and on overall alcohol use. In the original countries these laws were not very much noticed, being relatively small incremental steps in the accretion of policy. However, over the years, the Scandinavian countries won a reputation (not completely earned, in our opinion) for having dealt successfully with drunk driving, and the laws were copied in other jurisdictions where they represented a sharp break with tradition and therefore were much more noticeable. The first important adoption of this model outside Scandinavia came in Great Britain in 1967. The Road Safety Act of that year prohibited driving or attempting to drive with a blood-alcohol concentration greater than .08 percent (the level that a man of medium build might reach after drinking four or five drinks on an empty stomach in the period of an hour). It was proposed that police be empowered to stop any driver and administer a screening breath test for blood alcohol (using a new device imported in mass quantities from West Germany, with much fanfare). Although this provision was rejected by Parliament, the final legislation permitted a test on anyone involved in an accident (regardless of fault) or committing a serious traffic law violation.

In addition to receiving official publicity, the British Road Safety Act was strengthened by media attention that continued for years as the complex law was challenged on numerous grounds by defendants seeking to escape the mandatory penalty of a year's license suspension. Evaluations of the law initially showed substantial deterrent effects: weekend night fatalities and serious injuries declined by more than half immediately following the imposition of the new rule, and attribution of the decline to the law was supported by the failure of comparable casualties to decline during non-drinking hours, as well as by behavioral data reported in polls and other sources (Ross, 1973; Saunders, 1975). However, the effect of the law was not permanent. Matters started to return to the *status quo ante* shortly after the effect was evident, and within about two years the effect could no longer be demonstrated.

Very much the same pattern has been noted in other countries. In 1978, a Scandinavian-type law was adopted in France, a country with one of the highest per capita levels of alcohol consumption in the world. Police were given permission to conduct road blocks in which all drivers passing through could be tested for alcohol, with predictable controversy and publicity. Again, weekend-night serious casualties declined significantly with the inception of the law, and again a reversion was immediately evident, bringing matters back to the *status quo ante* within a year (Ross et al., 1982).

In the Netherlands, a Scandinavian-type law introduced in 1974 was evaluated by the superior method of roadside surveys of blood-alcohol concentrations among random samples of drivers, which found an important decline coinciding with the law's inception, along with a trend toward prior levels (Noordzij, 1980). The coincidence of the Dutch law with the fuel crisis of the 1970s unfortunately compromises the probative value of this example, as does failure to find confirmation of the effect in data on total crash-related fatalities.

A Canadian law of 1969 was found in two independent studies (Carr et al., 1975; Chambers et al., 1976) to be effective in reducing crash-related fatalities, despite the fact that it was much less threatening than the others mentioned (police needed to have reasonable suspicion of an offense before requiring the test); and there appears to have been much less resistance to and hence notoriety for the Canadian law. As in the other instances, the effect of the Canadian law was not found to be long-lasting.

Two negative reports have been made concerning Scandinavian-type laws in Australia (Birrell, 1975) and New Zealand (Hurst, 1978), but both were methodologically deficient studies of what appear to have been relatively unpublicized innovations.

Enforcement Crackdowns Police activity to enforce drunk-driving laws is relevant to the deterrence prediction that threatened behavior will be reduced when punishment appears to be more certain. Perhaps the classic case occurred in 1975 when the chief constable of the county of Cheshire, England, experimentally required his men to demand breath tests for alcohol in all situations where the law permitted the request. This experiment was ''discovered'' by the local press, which elevated the police activity to the status of a campaign. The consequence was significant decreases in serious injuries and fatalities during the month in which the increased enforcement was maintained (Ross, 1977). Similar results were found in both Australia (Cameron et al., 1980) and New Zealand (Hurst and Wright, 1980), where despite the initial negative results reported for their Scandinavian-type laws, campaigns of enforcement for these laws were accompanied by impressive declines in casualties in the affected jurisdictions.

Among the most ambitious enforcement efforts were the U.S. Alcohol Safety Action Projects (ASAPs) funded by the United States Department of Transportation in 35 sites in the mid-1970s. It is estimated that more than $200 million in public funds was spent on these projects, which centered on increasing patrol as well as on streamlining the processing of the accused in the criminal justice system. Unfortunately, very much like the High-Impact Anti-Crime program for street crimes, the structuring of the ASAPs varied by city, and the local evaluations were on the whole incompetent. However, a final evaluation by competent if perhaps not disinterested U.S. Department of Transportation staff (1979) did find evidence for a deterrent effect, as measured by greater reductions in nighttime than in daytime fatal crashes, in 12 of the 25 sites, and in 8 of the 13 sites where the absolute level of nighttime crashes and a moderate population growth rate rendered evaluation less problematic.

Severity of Punishment

Although efforts to increase the severity of punishment for drunk drivers have probably been much more frequent than those directed at certainty, there are few published evaluations. The efforts have taken the form of increasing statutory penalties and of judicial crackdowns increasing the actual penalties for drunk driving in various jurisdictions.

Statutory Changes Many statutory changes in the penalty for drunk driving have been accomplished as part of broad packages of countermeasures, some of which also relate to increasing certainty. One example of a law that appears to have been directed only toward increasing the perceived severity of penalties took place in Finland in 1950, when the maximum sentence for drunk driving was doubled from two to four years, with the provision for six years in the event of serious bodily injury resulting from the offense, and seven years for causing death. Although there was a decline in crash-related fatalities in subsequent years, it proved not to be possible to attribute this decline to the law because it was greater for less serious crashes than for fatal crashes, whereas the latter are more likely to involve drunk driving. Furthermore, the drop was greater for multiple-vehicle crashes than for single-vehicle crashes, the latter again more likely to involve alcohol (Ross, 1975).

Judicial Crackdowns In Chicago in 1970 the supervising judge of the traffic court decreed that all defendants judged guilty of drunk driving during claims of success for this campaign, careful analysis of the data found that the decline could not be distinguished from chance variation. Furthermore, data from Milwaukee, chosen as a comparison jurisdiction, showed an even greater proportional decline, although as in Chicago it was not statistically significant (Robertson et al., 1973).

Similar findings are reported from a city in New South Wales, Australia, where a local magistrate declared his intention to increase greatly the penalties for drunk drivers. Research showed that serious crashes did not decline perceptibly even though, unlike in Chicago, the threatened penalties were actually put into effect in most cases (Misner and Ward, 1975).

There is evidence in these studies that the criminal justice system reacts in a way that vitiates the declared severity of actual punishments. For example, in Chicago, it was found that convictions declined where drivers were accused in the absence of chemical tests in evidence. In a more recent study of jurisdictions adopting mandatory jail sentences for first offender DUI (driving under the influence) defendants, Gropper et al. (1983) report that there were important increases in not-guilty pleas, in jury trials, and in failures to appear for trial as well as dismissals and not-guilty findings. Moreover, the eventual punishment for those nonetheless convicted was slowed by considerable increases in delay between arrest and conviction.

Summary

The research to date on attempts to deter drunk drivers suggests that measures directed at increasing the perceived certainty of punishment can have a sharp, immediate deterrent effect on the proscribed behavior. Rates of crashes likely to involve alcohol decline sharply at the inception of well-publicized laws that simplify apprehension and prosecution, and during well-publicized campaigns of police enforcement. The extent of the observed declines in crashes is impressive in light of the fact that crashes involve many causal factors other than alcohol. Deterrent effects have been found in virtually all well-designed studies of significant interventions, in many countries throughout the world. However, these effects universally disappear over time, a matter of several months or a few years at most. One possible explanation for this fact is that the very low levels of actual likelihood of punishment are insufficient to continue an initial impression of reasonable certainty of punishment for the violator.

No deterrent effect is evident for legal interventions that are directed solely at increasing the severity of punishment. Although there are only a few reported studies supporting this generalization, there are no negative findings. It seems plausible to attribute this disconfirmation of deterrent expectations to the very low risk of punishment of any kind, which permits the violator to regard the threat as negligible, and to possible public perception of the fact that the criminal justice system does not necessarily deal any more severely after those interventions than before. Severity-based interventions are found to produce undesired and unanticipated side effects through the discretion of legal actors, behavior that may reflect the sense that actual offenses detected are a haphazard selection from a much larger population of undetected offenses, so that those charged are in a meaningful sense unlucky.

RESEARCH AND POLICY IMPLICATIONS

We must conclude that, as tests of the scientific validity of the deterrence proposition, existing research is inadequate. For many years social science research simply ignored the issue of deterrence. More recently the size and scope of the deterrence literature has increased dramatically, but it is still incapable of resolving basic theoretical questions, for several reasons.

First, evaluations of the deterrent effects of punishment are often based on changes in formal laws rather than changes in actual enforcement behaviors. This issue is particularly important with regard to sanction severity. Proclaimed increases in the severity of penalties repeatedly have been found to be vitiated by the reluctance or incapacity of legal agents to actually implement the new penalties. As for certainty, most interventions may be described as having increased the objective probability of punishment from "negligible" to "trivial" levels. The sheer resistance of the criminal justice system to piecemeal implementation of new penal sanctions may be the major finding of studies ostensibly testing the deterrence proposition.

Second, most prior research on deterrence relies on the unsupported assumption that changes in objective levels of certainty and severity of punishment are reflected in the *perceptions* that are the subject of the theoretical proposition being tested. Where the risk levels are extremely low, as is common in the situations being studied here, and where actual punishments are not increased despite policymakers' intentions, it is hazardous to assume that perceived certainty and severity of punishment have been changed.

Third, and more serious for studies of street crime than of drunk driving, is the general inadequacy of research design. Much of the research on street crime is based on correlational and econometric analyses, the defects of which have been well exposed (e.g., Greenberg, 1977; Blumstein et al., 1978). In the few classical experiments, control groups are often contaminated by proximity to experimental groups. Reliability and validity of measurement are serious problems for most of the studies of street crime interventions, and many of the time-series quasi experiments fail to control for the possibility that events other than the deterrence-based legal interventions may have caused declines in violations. For example, interventions against drunk driving in recent years have frequently coincided with economic transitions or crises in fuel availability.

But despite the admitted weaknesses of evidence in individual studies, the accumulated literature supports a number of tentative conclusions. The deterrence proposition is generally supported in evaluations of legal interventions bearing on the certainty of punishment for the offender. In the short run, at least, there is clear evidence that offense rates decline. In the long run, however, matters are not so clear, very likely because the deterrent effects depend on an overestimation of the chances of apprehension by the relevant public due to publicity and media attention surrounding the interventions. This impression may be difficult to maintain in the face of daily experience that fails to support it.

In contrast, there is very little evidence favoring deterrence in the matter of severity of punishment, even in the short run. One explanation is that the relevant public may readily learn or come to expect that the declared severity of penalties is compromised by resistance to change on the part of legal actors. An even more appealing explanation lies in a possible interaction between perceived severity and certainty: where the likelihood of any punishment is very low, as it very often seems to be, the prospective offender discounts even severe penalties as negligible.

Accepting this statement of the evidence, several questions can be raised for policy considerations. First, why does the deterrence approach, especially in the matter of severe punishment, remain so popular as a basis for countermeasures against these and other social problems? Second, what are the prospects for obtaining long-term deterrent results through increasing the probabilities of punishment? Third, what alternatives might be proposed as a basis for more rational countermeasures?

The Continued Reliance on Deterrence

One reason for the continued tendency to invoke deterrence-related countermeasures may be the intuitive appeal of the deterrence proposition. Introspection informs us that we often refrain from prohibited acts because of threatened punishment. Moreover, our daily experience in the marketplace confirms an economic counterpart of the deterrence proposition, which is that when the price of any good is raised, everything else remaining the same, less is consumed. If this intuitive confirmation fails us in nonmarket circumstances, it may be because much criminal and other socially problematic behavior faces threats with unusually small probabilities of fulfillment, in which rational calculation and behavior are noted to be uncertain.

The utilities involved in deciding whether to drink and drive, or to steal from despised or helpless victims, may be more like those associated with gambling, where hordes of people enjoy participation in mathematically unfair games, than like ordinary market behavior.

Second and, we think, very important, the cost of deterrence-based countermeasures tends to be delayed and obscure, while the costs of alternatives may be daunting. Increased severity, in particular, can seemingly be invoked with the stroke of a pen. In fact, as the grossly overcrowded conditions in many of the nation's prisons now attest, increasing the severity of punishment may be more costly in practice than it appears when considered as a basis for policy.

Third, deterrence-based measures frequently relieve established institutions and vested interests that might have much to lose from other countermeasures. The appeal of deterrence-based approaches to drunk driving, for instance, rests heavily on the assumption that the problem lies with a small minority of irresponsible deviants. Neither the alcoholic beverage nor the automobile industry bears responsibility in this conception of the problem. Likewise, massive public expenditures aimed at deterring street crime, though ineffective, focus attention on individual deviants and away from possible defects in the social structure, including gross inequalities in life chances among different perpetrator groups.

Fourth, deterrence-based reasoning may provide a cover for retribution-based motives fueling popular movements. The person injured by a mugger or drunk driver understandably may feel that muggers and drunk drivers deserve punishment, but demands for action based on this feeling may be more successful when legitimated by the promise of reductions in future damage from these causes. Perhaps in part for this reason, the anti-drunk-driving movement has been more insistent on enacting severe penalties than on requiring occupant-protection devices that would be much more effective in reducing fatalities.

The Prospect for Increased Certainty

Deterrence-based policy seems to founder on the low actual rates of apprehension and punishment for offenders. It is possible that greatly increased investments in criminal justice might in the end be effective; nothing in the research literature refutes this possibility. But we doubt that this is a profitable line of endeavor. There is little in accumulated experience with street crimes or drunk driving to suggest where, in the scale of probability of punishment, a threshold of appreciably greater effectiveness may lie, but it seems likely to involve levels of police intrusiveness and expense otherwise unknown in democratic societies. The American public is deeply ambivalent about this approach and fundamentally negative when presented with unbiased estimates of the costs. For example, although the San Diego study showed that field interrogation by police might be an effective crime deterrent, many commentators have rejected this kind of police activity. Charles Reich (1979:113) has put it thus: "I fully recognize that safety is important and that safety requires measures. But other qualities also require measures: I mean independence, boldness, creativity, high spirits." Many Americans envy the tranquility of Japanese urban society, but most would find the Japanese system of policing, in which neighborhood police provide their headquarters with updated information on the daily lives of residents of their jurisdiction, too intrusive.

The research literature also indicates that legal actors resist substantive changes in their job-related behavior. This may be caused by feelings of injustice when the prescribed penalty fails to fit the crime, or it may simply be rationalized human laziness or procedural resistance to change. As Zimring (1978:171) states the case, "The resiliency of courts and police when policy changes are induced by outside money investments is formidable." Thus, in a recent study of the response of police to the mandate to "do something about rape" in a large midwestern city (LaFree, 1981), it was clear that the police changed those things easiest to change—primarily recordkeeping—while doing little about the things that the deterrence proposition suggests would be most important for actually reducing rape, such as increasing arrests and filing more felony complaints.

Alternatives to Deterrence

The social science literature on both street crime and drunk driving suggests then it may be more fruitful to consider offenders as rational reactors to their social environments rather than as irrational, maladapted, or pathological deviants. The case may be easier to make for the general population in such matters as white-collar crime (e.g., filing deceptive tax returns) than among street criminals, but the literature suggests that when the world is viewed from the deviant's position in society much of the problematical behavior appears normal and predictable (cf. Lempert, 1981–1982, for a discussion of this point among fathers ordered to pay child support). Surely the case is easily made for the drunk driver, who exists in a society that institutionalizes the use of alcohol as a social lubricant and mandates dependence on the private automobile for most of its members. Because the probability of the most severe consequence is in reality minuscule (one fatality for about one third of a million miles of drunk driving), For both street crime and drunk driving, this line of thought has led to suggestions for modifying the situation, rather than the individual, to reduce the problematic behavior. Street crimes may perhaps be better reduced by "hardening the target" than by deterring or, for that matter, reforming potential criminals. For example, Oscar Newman (1972) has written extensively on how urban design can reduce crime by making living areas more "defensible." Similarly, drunk driving has been reduced in the 18- to 21-year-old age group as a consequence of raising the drinking age in various states (Wagenaar, 1982). Effects are also being found for the restriction of young drivers to daylight (nondrinking-hour) driving only (Preusser et al., 1984). There is evidence for the view that raising the price of alcoholic beverages through taxation would reduce drunk driving along with a host of other alcohol-related problems (Moore and Gerstein, 1981). Much better public transportation, including subsidized taxi-like service, is another promising countermeasure.

A somewhat different approach to these problems is based on accepting the difficulty of fundamental changes in social institutions as well as behavior, and striving to make the results of the problematic behavior less damaging to the victims. For example, drunk driving would be of relatively less consequence if it did not entail deaths and injuries. By "padding" the car with passive restraints and the highway with soft shoulders and yielding barriers around fixed obstacles, the inevitable crashes caused by alcohol (and a host of other factors) could be better absorbed by society. Similarly, the cost of such activities as burglary could be lowered, though not eliminated, by social insurance schemes to repay the victims' financial losses.

Perhaps the main reason these types of countermeasures are less attractive than those based on deterrence is that their costs are out front, and they are not trivial. We further admit that they are unlikely to be "solutions" to the problems they address; most are merely mitigating. However, it strikes us as a more rational policy to experiment along these lines in hopes of finding economically sound mitigants than to follow the chimera of deterrence-based solutions that experience repeatedly shows to be inadequate in the context for which they are proposed.

* * *

We would like to acknowledge the helpful comments of Gwynn Nettler and Jack Gibbs.

REFERENCES

American Friends Service Committee
1971 *Struggle for Justice: A Report on Crime and Punishment in America.* New York: Hill and Wang.

ndenaes, J.

1974 *Punishment and Deterrence*. Ann Arbor: University of Michigan Press.

rcher, D., Gartner, R., and Beittel, M.

1983 Homicide and the death penalty: a cross-national test of a deterrence hypothesis. *Journal of Criminal Law and Criminology* 74:991–1013.

rrell, J.

1975 The compulsory breathaliser .05 percent legislation in Victoria. Pp. 775–785 in S. Israelstam and S. Lambert, eds., *Alcohol, Drugs and Traffic Safety*. Toronto: Addiction Research Foundation of Ontario.

umstein, A., Cohen, J., and Nagin, D., eds.

1978 *Deterrence and Incapacitation: Estimating the Effects of Criminal Sanctions on Crime Rates*. Panel on Research on Deterrent and Incapacitative Effects, National Research Council. Washington, D.C.: National Academy of Sciences.

owers, W., and Pierce, G.

1975 The illusion of deterrence in Isaac Ehrlich's research on capital punishment. *Yale Law Journal* 85:187–208.

oydstun, J.

1975 *San Diego Field Interrogation: Final Report*. Washington, D.C.: The Police Foundation.

ameron, M., Strang, P., and Vulcan, A.

1980 Evaluation of Random Breath Testing in Victoria, Australia. Paper presented at the Eighth International Conference on Alcohol, Drugs and Traffic Safety, Stockholm.

ampion, D.R.

1955 Does the death penalty protect state police? Appendix F, Part I, Minutes of Proceedings and Evidence, No. 20, Joint Committee of the Senate and House of Commons on Capital Punishment and Corporal Punishment and Lotteries, Canadian Parliament. Reported in part in H. Bedau, ed., *The Death Penalty in America*, revised ed. Garden City: Anchor Books.

arr, B., Goldberg, H., and Farbar, C.

1975 The Canadian breathaliser legislation: an inferential evaluation. Pp. 679–687 in S. Israelstam and S. Lambert, eds., *Alcohol, Drugs and Traffic Safety*. Toronto: Addiction Research Foundation of Ontario.

naiken, J.

1976 What's known about deterrent effects of police activities. *The Rand Paper Series*. Santa Monica: Rand Corporation.

naiken, J., Lawless, M., and Stevenson, R.

1974 *The Impact of Police Activity on Crime: Robberies on the New York Subway System*. Report number R-1424-NYC. Santa Monica: Rand Corporation.

nambers, L., Roberts, R., and Voeller, C.

1976 The epidemiology of traffic accidents and the effect of the 1969 breathaliser amendment in Canada. *Accident Analysis and Prevention* 8:201–206.

nelimsky, E.

1976 *High Impact Anti-Crime Program*. National Institute of Law Enforcement and Criminal Justice, Law Enforcement Assistance Administration: U.S. Department of Justice.

niricos, T., and Waldo, G.

1970 Punishment and crime: an examination of some empirical evidence. *Social Problems* 18:200–217.

ook, P.

1977 Punishment and crime: a critique of current findings concerning the preventive effects of punishment. *Law and Contemporary Problems* 41:164–204.

Ehrlich, I.
1975 The deterrent effect of capital punishment: a question of life and death. *The American Economic Review* 65:397–417.

Forst, B.
1976 Participation in illegitimate activities: further empirical findings. *Policy Analysis* 2:477–492. .

Gallagher, F.
1978 Appendix: an annotated bibliography of deterrence evaluations, 1970–1975. Pp. 174–186 in A. Blumstein, J. Cohen, and D. Nagin, eds., *Deterrence and Incapacitation: Estimating the Effects of Criminal Sanctions on Crime Rates*. Washington, D.C.: National Academy of Sciences.

Gibbons, D.
1979 *The Criminological Enterprise: Theories and Perspectives*. Englewood Cliffs, N.J.: Prentice-Hall.

Gibbs, J.
1975 *Crime, Punishment and Deterrence*. New York: Elsevier.

Graves, W.
1956 The deterrent effect of capital punishment in California. Reprinted in part in H. Bedau, ed., *The Death Penalty in America*, revised ed. Garden City: Anchor Books.

Greenberg, D.
1977 Deterrence research and social policy. Pp. 281–295 in S. Nagel, ed., *Modeling the Criminal Justice System*. Beverly Hills: Sage.

Gropper, B., Martorama, C., Mock, L., O'Connor, M., and Travers, W.
1983 *The Impacts of Mandatory Confinement for Drunk Driving on Criminal Justice Operations* (summary report). Technical report. Washington, D.C.: U.S. Department of Justice, National Institute of Justice.

Hindelang, M.
1976 *Criminal Victimization in Eight American Cities: A Descriptive Analysis of Common Theft and Assault*. Cambridge: Ballinger.

Hurst, P.
1978 Blood test legislation in New Zealand. *Accident Analysis and Prevention* 10:287–296.

Hurst, P., and Wright, P.
1980 Deterrence at Last: The Ministry of Transport's Alcohol Blitzes. Paper presented at the Eighth International Conference on Alcohol, Drugs and Traffic Safety, Stockholm.

Jeffery, C.
1972 The historical development of criminology. Pp. 458–498 in H. Mannheim, ed., *Pioneers in Criminology*, 2nd ed. Montclair, N.J.: Patterson Smith.

Kelling, G., and Pate, A.
1974 *The Kansas City Preventive Patrol Experiment*. Washington, D.C.: The Police Foundation.

Klein, L., Forst, B., and Filotov, V.
1978 The deterrent effect of capital punishment: an assessment of the estimates. Pp. 336–360 in A. Blumstein, J. Cohen, and D. Nagin, eds., *Deterrence and Incapacitation: Estimating the Effects of Criminal Sanctions on Crime Rates*. Washington, D.C.: National Academy of Sciences.

LaFree, G.
1981 Official reactions to social problems: police decisions in sexual assault cases. *Social Problems* 28:582–594.

Lempert, R.
1981–1982 Organizing for deterrence: lessons from a study of child support. *Law and Society Review* 16:514–568.

Lipton, D., Martinson, R., and Wilkes, J.
1975 *The Effectiveness of Correctional Treatment: A Survey of Treatment Evaluation Studies*. New York: Holt, Rinehart, and Winston.

Martinson, R.
1974 What works? Questions and answers about prison reform. *The Public Interest* 35:22–34.

Misner, R., and Ward, P.
1975 Severe penalties for driving offenses: deterrence analysis. *Arizona State Law Journal* 1975:677–713.

Moore, M., and Gerstein, D.
1981 *Alcohol and Public Policy: Beyond the Shadow of Prohibition*. Washington, D.C.: National Academy Press.

Nagin, D.
1978 General deterrence: a review of the empirical evidence. Pp. 95–139 in A. Blumstein, J. Cohen, and D. Nagin, eds., *Deterrence and Incapacitation: Estimating and Effects of Criminal Sanctions on Crime Rates*. Washington, D.C.: National Academy of Sciences.

Newman, O.
1972 *Defensible Space*. New York: Macmillan.

Noordzij, P.
1980 Recent Trends in Countermeasures and Research on Drinking and Driving in the Netherlands. Paper presented at the Eighth International Conference on Alcohol, Drugs and Traffic Safety, Stockholm.

President's Commission on Law Enforcement and Administration of Justice
1967 *The Challenge of Crime in a Free Society*. Washington, D.C.: U.S. Government Printing Office.

Press, S.
1971 *Some Effects of an Increase in Police Manpower in the 20th Precinct of New York City*. Santa Monica: Rand Corporation.

Preusser, D., Williams, A., Zador, P., and Blomberg, R.
1984 The effect of curfew laws on motor vehicle crashes. *Law and Policy*.

Reckless, W.
1967 *The Crime Problem*, 4th ed. Englewood Cliffs, N.J.: Prentice-Hall.

Reich, C.
1979 Police questioning of law abiding citizens. Pp. 107–113 in J. Bonsignore, E. Katsh, P. d'Errico, R. Pipkin, S. Arons, and J. Rifkin, eds., *Before the Law*. Boston: Houghton Mifflin.

Robertson, L., Rich, R., and Ross, H.
1973 Jail sentences for driving while intoxicated: a judicial policy that failed. *Law and Society Review* 8:55–67.

Ross, H.
1973 Law, science and accidents: the British road safety act of 1967. *Journal of Legal Studies* 2:1–78.

1975 The Scandinavian myth: the effectiveness of drinking-driving legislation in Sweden and Norway. *Journal of Legal Studies* 4:285–310.

1977 Deterrence regained: the Cheshire constabulary's "breathaliser blitz." *Journal of Legal Studies* 6:241–249.

Ross, H., McCleary, R., and Epperlein, T.
1982 Deterrence of drinking and driving in France: an evaluation of the law of July 12, 1978. *Law and Society Review* 16:345–374.

Saunders, A.

1975 Seven years experience of blood-alcohol limits in Britain. Pp. 845–853 in S. Israelstam and S. Lambert, eds., *Alcohol, Drugs and Traffic Safety*. Toronto: Addiction Research Foundation of Ontario.

Savitz, L.

1958 A study in capital punishment. *Journal of Criminal Law, Criminology and Police Science* 49:338–341.

Schwartz, B.

1968 The effect in Philadelphia of Pennsylvania's increased penalties for rape and attempted rape. *Journal of Criminal Law, Criminology and Police Science* 59:509–515.

Scull, A.

1977 *Decarceration*. Englewood Cliffs, N.J.: Prentice-Hall.

Sellin, T.

1959 *The Death Penalty*. Philadelphia: American Law Institute.

Sutherland, E.

1924 *Criminology*. Philadelphia: Lippincott.

Tittle, C.

1980 *Sanctions and Social Deviance*. New York: Praeger.

U.S. Department of Transportation

1979 *Alcohol Safety Action Projects: Evaluation of Operations, Data, Tables of Results and Formulation*. Washington, D.C.: National Highway Traffic Safety Administration.

Vold, G.

1979 *Theoretical Criminology*, 2nd ed. New York: Oxford University Press.

Wagenaar, A.

1982 Effects of the minimum drinking age on automotive crashes involving young drivers. *UMTRI Research Review* 13:3–12.

Walker, N.

1969 *Sentencing in Rational Society*. London: Penguin Press.

Wilson, J.

1975 *Thinking About Crime*. New York: Basic Books.

1983a Thinking about crime: the debate over deterrence. *The Atlantic Monthly*. September, Pp. 72–88.

1983b Crime and American culture. *The Public Interest* 1983:22–48.

Zeisel, H.

1976 The deterrent effect of the death penalty: facts vs. faiths. In P. Kurland, ed., *The Supreme Court Review: 1976*. Chicago: University of Chicago Press.

Zimring, F.

1978 Policy experiments in general deterrence: 1970–1975. Pp. 140–174 in A. Blumstein, J. Cohen, and D. Nagin, eds., *Deterrence and Incapacitation: Estimating the Effects of Criminal Sanctions on Crime Rates*. Washington, D.C.: National Academy of Sciences.

Zimring, F., and Hawkins, G.

1973 *Deterrence*. Chicago: University of Chicago Press.

Choices, Values, and Frames

DANIEL KAHNEMAN and AMOS TVERSKY

The making of decisions is perhaps the most fundamental activity that characterizes living creatures. Consequently, the attempt to understand, explain, and predict individual choice behavior has been a major goal of the behavioral and social sciences. Indeed, economics, psychology, sociology, and political science are all concerned with the analysis of decisions made by consumers, patients, voters, and politicians. The study of decisions addresses both normative and descriptive questions. The normative analysis is concerned with the nature of rationality and the logic of decisionmaking. The descriptive analysis, in contrast, is concerned with people's beliefs and preferences as they are, not as they should be. The tension between normative and descriptive considerations characterizes much of the study of judgment and choice.

Analyses of decisionmaking commonly distinguish risky and riskless choices. The paradigmatic example of decision under risk is the acceptability of a gamble that yields monetary outcomes with specified probabilities. A typical riskless decision concerns the acceptability of a transaction in which a good or a service is exchanged for money or labor. In the first part of this article we present an analysis of the cognitive and psychophysical factors that determine the value of risky prospects. In the second part we extend this analysis to transactions and trades.

This is a slightly modified version of a paper that appeared in the *American Psychologist* on April 1984, vol. 39, no. 4, pp. 341–350. Copyright 1984 by the American Psychological Association. Adapted by permission of the publisher and authors. This work was supported by grant NR 197-058 from the Office of Naval Research.

RISKY CHOICE

The making of decisions is commonly complicated by the presence of uncertainty or risk. In general we cannot predict with certainty tomorrow's weather, the outcome of a medical treatment, or the future value of gold. Hence the decisions whether to undergo surgery, carry an umbrella, or buy gold must be made without advance knowledge of their consequences. It is therefore natural that the study of decisionmaking under risk has focused on choices between simple gambles with monetary outcomes and specified probabilities in the hope that these simple problems will reveal basic attitudes toward risk and value.

We shall describe an approach to the analysis of risky choice that derives many of its hypotheses from a psychophysical analysis of value and probability. Psychophysics is the study of the relations between physical magnitudes, such as length or money, and their psychological counterparts, such as perceived length or utility.

The psychophysical approach to decisionmaking can be traced to a remarkable essay that Daniel Bernoulli published in 1738 (Bernoulli, 1738/1954) in which he attempted to explain why people are generally averse to risk and why risk aversion decreases with increasing wealth. To illustrate risk aversion and Bernoulli's analysis, consider the choice between a prospect that offers an 85 percent chance to win \$1,000 (with a 15 percent chance to win nothing) and the alternative of receiving \$800 for sure. A large majority of people prefer the sure thing over the gamble, although the gamble has higher (mathematical) expectation. The expectation of a monetary gamble is a weighted average, where each possible outcome is weighted by its probability of occurrence. The expectation of the gamble in this example is .85 × \$1,000 + .15 × \$0 = \$850, which exceeds the expectation of \$800 associated with the sure thing. The preference for the sure gain is an instance of risk aversion. In general, a preference for a sure outcome over a gamble that has higher or equal expectation is called risk averse, and the rejection of a sure thing in favor of a gamble of lower or equal expectation is called risk seeking.

Bernoulli suggested that people do not evaluate prospects by the expectation of their monetary outcomes, but rather by the expectation of the subjective value of these outcomes. The subjective value of a gamble is again a weighted average, but now it is the subjective value of each outcome that is weighted by its probability. To explain risk aversion within this framework, Bernoulli proposed that subjective value, or utility, is a concave function of money. In such a function, the difference between the utilities of \$200 and \$100, for example, is greater than the utility difference between \$1,200 and \$1,100. It follows from concavity that the subjective value attached to a gain of $800 is more than 80 percent of the value of a gain of $1,000. Consequently, the concavity of the utility function entails a risk averse preference for a sure gain of $800 over an 80 percent chance to win $1,000, although the two prospects have the same monetary expectation.

It is customary in decision analysis to describe the outcomes of decisions in terms of total wealth. For example, an offer to bet $20 on the toss of a fair coin is represented as a choice between an individual's current wealth W and an even chance to move to W + $20 or to W − $20. This representation appears psychologically unrealistic: People do not normally think of relatively small outcomes in terms of states of wealth but rather in terms of gains, losses, and neutral outcomes (such as the maintenance of the status quo). If the effective carriers of subjective value are changes of wealth rather than ultimate states of wealth, as we propose, the psychophysical analysis of outcomes should be applied to gains and losses rather than to total assets. This assumption plays a central role in a treatment or risky choice that we called prospect theory (Kahneman and Tversky, 1979). Introspection as well as psychophysical measurements suggest that subjective value is a concave function of the size of a gain. The same generalization applies to losses as well. The difference in subjective value between a loss of $200 and a loss of $100 appears greater than the difference in subjective value between a loss of $1,200 and a loss of $1,100. When the value functions for gains and for losses are pieced together, we obtain an S-shaped function of the type displayed in Figure 1.

The value function shown in Figure 1 is (a) defined on gains and losses rather than on total wealth, (b) concave in the domain of gains and convex in the domain of losses, and (c) considerably steeper for losses than for gains. The last property, which we label *loss aversion*, expresses the intuition that a loss of $X is more aversive than a gain of $X is attractive.

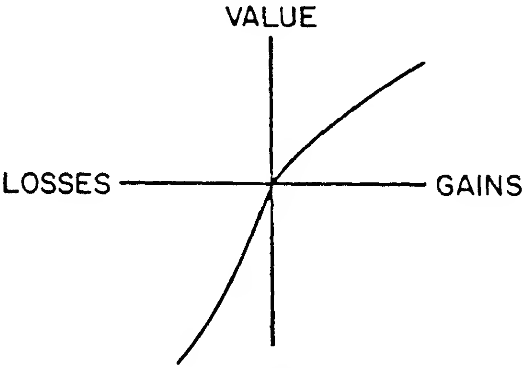


FIGURE 1. A hypothetical value function.

Loss aversion explains people's reluctance to bet on a fair coin for equal stakes: The attractiveness of the possible gain is not nearly sufficient to compensate for the aversiveness of the possible loss. For example, most respondents in a sample of undergraduates refused to stake $10 on the toss of a coin if they stood to win less than $30.

The assumption of risk aversion has played a central role in economic theory. However, just as the concavity of the value of gains entails risk aversion, the convexity of the value of losses entails risk seeking. Indeed, risk seeking in losses is a robust effect, particularly when the probabilities of loss are substantial. Consider, for example, a situation in which an individual is forced to choose between an 85 percent chance to lose $1,000 (with a 15 percent chance to lose nothing) and a sure loss of $800. A large majority of people express a preference for the gamble over the sure loss. This is a risk-seeking choice because the expectation of the gamble (−$850) is inferior to the expectation of the sure loss (−$800). Risk seeking in the domain of losses has been confirmed by several investigators (Fishburn and Kochenberger, 1979; Hershey and Schoemaker, 1980; Payne et al., 1980; Slovic et al., 1982). It has also been observed with nonmonetary outcomes, such as hours of pain (Eraker and Sox, 1981) and loss of human lives (Fischhoff, 1983; Tversky, 1977; Tversky and Kahneman, 1981). Is it wrong to be risk averse in the domain of gains and risk seeking in the domain of losses? These preferences conform to compelling intuitions about the subjective value of gains and losses, and the presumption is that people should be entitled to their own values. However, we shall see that an S-shaped value function has implications that are normatively unacceptable.

To address the normative issue we turn from psychology to decision theory. Modern decision theory can be said to begin with the pioneering work of von Neumann and Morgenstern (1947), who laid down several qualitative principles, or axioms, that should govern the preferences of a rational decisionmaker. Their axioms included transivity (if A is preferred to B and B is preferred to C, then A is preferred to C), and substitution (if A is preferred to B, then an even chance to get A or C is preferred to an even chance to get B or C), along with other conditions of a more technical nature. The normative and the descriptive status of the axioms of rational choice have been the subject of extensive discussions. In particular, there is convincing evidence that people do not always obey the substitution axiom, and considerable disagreement exists about the normative merit of this axiom (e.g., Allais and Hagen, 1979). However, all analyses of rational choice incorporate two principles: *dominance* and *invariance*. Dominance demands that if prospect A is at least as good as prospect B in every respect and better than B in at least one respect, then A should be preferred to B. Invariance requires that the preference order between prospects should not depend on the manner in which they are described. In particular, two versions of a choice problem that are recognized to be equivalent when shown together should elicit the same preference even when shown separately. We now show that the requirement of invariance, however elementary and innocuous it may seem, cannot generally be satisfied.

Framing of Outcomes

Risky prospects are characterized by their possible outcomes and by the probabilities of these outcomes. The same option, however, can be framed or described in different ways (Tversky and Kahneman, 1981). For example, the possible outcomes of a gamble can be framed either as gains and losses relative to the status quo or as asset positions that incorporate initial wealth. Invariance requires that such changes in the description of outcomes should not alter the preference order. The following pair of problems illustrates a violation of this requirement. The total number of respondents in each problem is denoted by *N*, and the percentage who chose each option is indicated in parentheses.

Problem 1 (*N* = 152): Imagine that the United States is preparing for the outbreak of an unusual Asian disease, which is expected to kill 600 people. Two alternative programs to combat the disease have been proposed. Assume that the exact scientific estimates of the consequences of the programs are as follows:

If Program A is adopted, 200 people will be saved. (72%)

If Program B is adopted, there is a one-third probability that 600 people will be saved and a two-thirds probability that no people will be saved. (28%)

Which of the two programs would you favor?

The formulation of Problem 1 implicitly adopts as a reference point a state of affairs in which the disease is allowed to take its toll of 600 lives. The outcomes of the programs include the reference state and two possible gains, measured by the number of lives saved. As expected, preferences are risk averse: A clear majority of respondents prefer saving 200 lives for sure over a gamble that offers a one-third chance of saving 600 lives. Now consider another problem in which the same cover story is followed by a different description of the prospects associated with the two programs:

Problem 2 (*N* = 155):

If Program C is adopted, 400 people will die. (22%)

If Program D is adopted, there is a one-third probability that nobody will die and a two-thirds probability that 600 people will die. (78%)

It is easy to verify that options C and D in Problem 2 are undistinguishable in real terms from options A and B in Problem 1, respectively. The second version, however, assumes a reference state in which no one dies of the disease. The best outcome is the maintenance of this state and the alternatives are losses measured by the number of people that will die of the disease. People who evaluate options in these terms are expected to show a risk seeking preference for the gamble (option D) over the sure loss of 400 lives. Indeed, there is more risk seeking in the second version of the problem than there is risk aversion in the first.

The failure of invariance is both pervasive and robust. It is as common among sophisticated respondents as among naive ones, and it is not eliminated even when the same respondents answer both questions within a few minutes. Respondents confronted with their conflicting answers are typically puzzled. Even after rereading the problems, they still wish to be risk averse in the ''lives saved'' version; they wish to be risk seeking in the ''lives lost'' version; and they also wish to obey invariance and give consistent answers in the two versions. In their stubborn appeal, framing effects resemble perceptual illusions more than computational errors.

The following pair of problems elicits preferences that violate the dominance requirement of rational choice.

Problem 3 (N = 86): Choose between:
E. 25% chance to win $240 and
75% chance to lose $760 (0%)
F. 25% chance to win $250 and
75% chance to lose $750 (100%)

It is easy to see that F dominates E. Indeed, all respondents chose accordingly.

Problem 4 (N = 150): Imagine that you face the following pair of concurrent decisions. First examine both decisions, then indicate the options you prefer.

Decision (i) Choose between:
A. a sure gain of $240 (84%)
B. 25% chance to gain $1,000 and
75% chance to gain nothing (16%)
Decision (ii) Choose between:
C. a sure loss of $750 (13%)
D. 75% chance to lose $1,000 and
25% chance to lose nothing (87%)

As expected from the previous analysis, a large majority of subjects made a risk-averse choice for the sure gain over the positive gamble in the first decision, and an even larger majority of subjects made a risk-seeking choice for the gamble over the sure loss in the second decision. In fact, 73 percent of the respondents chose A and D and only 3 percent chose B and C. The same pattern of results was observed in a modified version of the problem, with reduced stakes, in which undergraduates selected gambles that they would actually play.

Because the subjects considered the two decisions in Problem 4 simultaneously, they expressed in effect a preference for A and D over B and C. The preferred conjunction, however, is actually dominated by the rejected one. Adding the sure gain of $240 (option A) to option D yields a 25 percent chance to win $240 and a 75 percent chance to lose $760. This is precisely option E in Problem 3. Similarly, adding the sure loss of $750 (option C) to option B yields a 25 percent chance to win $250 and a 75 percent chance to lose $750. This is precisely option F in Problem 3. Thus, the susceptibility to framing and the S-shaped value function produce a violation of dominance in a set of concurrent decisions.

The moral of these results is disturbing: Invariance is normatively essential, intuitively compelling, and psychologically unfeasible. Indeed, we conceive only two ways of guaranteeing invariance. The first is to adopt a procedure that will transform equivalent versions of any problem into the same canonical representation. This is the rationale for the standard admonition to students of business, that they should consider each decision problem in terms of total assets rather than in terms of gains or losses (Schlaifer, 1959). Such a representation would avoid the violations of invariance illustrated in the previous problems, but the advice is easier to give than to follow. Except in the context of possible ruin, it is more natural to consider financial outcomes as gains and losses rather than as states of wealth. Furthermore, a canonical representation of risky prospects requires a compounding of all outcomes of concurrent decisions (e.g., Problem 4) that exceeds the capabilities of intuitive computation even in simple problems. Achieving a canonical representation is even more difficult in other contexts such as safety, health, or quality of life. Should we advise people to evaluate the consequence of a public health policy (e.g., Problems 1 and 2) in terms of overall mortality, mortality due to diseases, or the number of deaths associated with the particular disease under study?

Another approach that could guarantee invariance is the evaluation of options in terms of their actuarial rather than their psychological consequences. The actuarial criterion has some appeal in the context of human lives, but it is clearly inadequate for financial choices, as has been generally recognized at least since Bernoulli, and it is entirely inapplicable to outcomes that lack an objective metric. We conclude that frame invariance cannot be expected to hold and that a sense of confidence in a particular choice does not ensure that the same choice would be made in another frame. It is therefore good practice to test the robustness of preferences by deliberate attempts to frame a decision problem in more than one way (Fischhoff et al., 1980).

The Psychophysics of Chances

Our discussion so far has assumed a Bernoullian expectation rule according to which the value, or utility, of an uncertain prospect is obtained by adding the utilities of the possible outcomes, each weighted by its probability. To examine this assumption, let us again consult psychophysical intuitions. Setting the value of the status quo at zero, imagine a cash gift, say of $300, and assign it a value of one. Now imagine that you are only given a ticket to a lottery that has a single prize of $300. How does the value of the ticket vary as a function of the probability of winning the prize? Barring utility for gambling, the value of such a prospect must vary between zero (when the chance of winning is nil) and one (when winning $300 is a certainty).

Intuition suggests that the value of the ticket is not a linear function of the probability of winning, as entailed by the expectation rule. In particular, an increase from 0 percent to 5 percent appears to have a larger effect than an increase from 30 percent to 35 percent, which also appears smaller than an increase from 95 percent to 100 percent. These considerations suggest a category-boundary effect: A change from impossibility to possibility or from possibility to certainty has a bigger impact than a comparable change in the middle of the scale. This hypothesis is incorporated into the curve displayed in Figure 2, which plots the weight attached to an event as a function of its stated numerical probability. The most salient feature of Figure 2 is that decision weights are regressive with respect to stated probabilities. Except near the endpoints, an increase of .05 in the probability of winning increases the value of the prospect by less than 5 percent of the value of the prize. We next investigate the implications of these psychophysical hypotheses for preferences among risky options.

In Figure 2, decision weights are lower than the corresponding probabilities over most of the range. Underweighting of moderate and high probabilities relative to sure things contributes to risk aversion in gains by reducing the attractiveness of positive gambles. The same effect also contributes to risk seeking in losses by attenuating the aversiveness of negative gambles. Low probabilities, however, are overweighted, and very low probabilities are either overweighted quite grossly or neglected altogether,

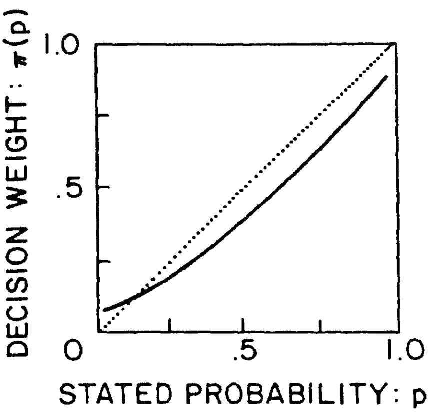


FIGURE 2. A hypothetical weighting function.

making the decision weights highly unstable in that region. The overweighting of low probabilities reverses the pattern described above: It enhances the value of long shots and amplifies the aversiveness of a small chance of a severe loss. Consequently, people are often risk seeking in dealing with improbable gains and risk averse in dealing with unlikely losses. Thus, the characteristics of decision weights contribute to the attractiveness of both lottery tickets and insurance policies.

The nonlinearity of decision weights inevitably leads to violations of invariance, as illustrated in the following pair of problems:

Problem 5 (N = 85): Consider the following two-stage game.

In the first stage, there is a 75% chance to end the game without winning anything and a 25% chance to move into the second stage. If you reach the second stage you have a choice between:

A. a sure win of $30	(74%)
B. 80% chance to win $45	(26%)

Your choice must be made before the game starts, i.e., before the outcome of the first stage is known. Please indicate the option you prefer.

Problem 6 (N = 81): Which of the following options do you prefer?

C. 25% chance to win $30	(42%)
D. 20% chance to win $45	(58%)

Because there is one chance in four to move into the second stage in Problem 5, prospect A offers a .25 probability of winning $30, and prospect B offers .25 × .80 = .20 probability of winning $45. Problems 5 and 6 are therefore identical in terms of probabilities and outcomes. However, the preferences are not the same in the two versions: A clear majority favors the higher chance to win the smaller amount in Problem 5, whereas the majority goes the other way in Problem 6. This violation of invariance has been confirmed with both real and hypothetical monetary payoffs (the present results are with real money), with human lives as outcomes, and with a nonsequential representation of the chance process.

We attribute the failure of invariance to the interaction of two factors: the framing of probabilities and the nonlinearity of decision weights. More specifically, we propose that in Problem 5 people ignore the first phase, which yields the same outcome regardless of the decision that is made, and focus their attention on what happens if they do reach the second stage of the game. In that case, of course, they face a sure gain if they choose option A and an 80 percent chance of winning if they prefer to gamble. Indeed, people's choices in the sequential version are practically identical to the choices they make between a sure gain of $30 and an 85 percent chance to win $45. Because a sure thing is overweighted in comparison with events of moderate or high probability (see Figure 2) the option that may lead to a gain of $30 is more attractive in the sequential version. We call this phenomenon the *pseudo-certainty* effect because an event that is actually uncertain is weighted as if it were certain.

A closely related phenomenon can be demonstrated at the low end of the probability range. Suppose you are undecided whether or not to purchase earthquake insurance because the premium is quite high. As you hesitate, your friendly insurance agent comes forth with an alternative offer: ''For half the regular premium you can be fully covered if the quake occurs on an odd day of the month. This is a good deal because for half the price you are covered for more than half the days.'' Why do most people find such probabilistic insurance distinctly unattractive? Figure 2 suggests an answer. Starting anywhere in the region of low probabilities, the impact on the decision weight of a reduction of probability from p to $p/2$ is considerably smaller than the effect of a reduction from $p/2$ to 0. Reducing the risk by half, then, is not worth half the premium.

The aversion to probabilistic insurance is significant for three reasons. First, it undermines the classical explanation of insurance in terms of a concave utility function. According to expected utility theory, probabilistic insurance should be definitely preferred to normal insurance when the latter is just acceptable (see Kahneman and Tversky, 1979). Second, probabilistic insurance represents many forms of protective action, such as having a medical checkup, buying new tires, or installing a burglar alarm system. Such actions typically reduce the probability of some hazard without eliminating it altogether. Third, the acceptability of insurance can be manipulated by the framing of the contingencies. An insurance policy that covers fire but not flood, for example, could be evaluated either as full protection against a specific risk, (e.g., fire) or as a reduction in the overall probability of property loss. Figure 2 suggests that people greatly undervalue a reduction in the probability of a hazard in comparison to the complete elimination of that hazard. Hence, insurance should appear more attractive when it is framed as the elimination of risk than when it is described as a reduction of risk. Indeed, Slovic, Fischhoff, and Lichtenstein (1982) showed that a hypothetical vaccine that reduces the probability of contracting a disease from 20 percent to 10 percent is less attractive if it is described as effective in half of the cases than if it is presented as fully effective against one of two exclusive and equally probable virus strains that produce identical symptoms.

Formulation Effects

So far we have discussed framing as a tool to demonstrate failures of invariance. We now turn attention to the processes that control the framing of outcomes and events. The public health problem illustrates a formulation effect in which a change of wording from "lives saved" to "lives lost" induced a marked shift in preference from risk aversion to risk seeking. Evidently, the subjects adopted the descriptions of the outcomes as given in the question and evaluated the outcomes accordingly as gains or losses. Another formulation effect was reported by McNeil, Pauker, Sox, and Tversky (1982). They found that preferences of physicians and patients between hypothetical therapies for lung cancer varied markedly when their probable outcomes were described in terms of mortality or survival. Surgery, unlike radiation therapy, entails a risk of death during treatment. As a consequence, the surgery option was relatively less attractive when the statistics of treatment outcomes were described in terms of mortality rather than in terms of survival.

A physician, and perhaps a presidential advisor as well, could influence the decision made by the patient or by the President, without distorting or suppressing information, merely by the framing of outcomes and contingencies. Formulation effects can occur fortuitously, without anyone being aware of the impact of the frame on the ultimate decision. They can also be exploited deliberately to manipulate the relative attractiveness of options. For example, Thaler (1980) noted that lobbyists for the credit card industry insisted that any price difference between cash and credit purchases be labeled a cash discount rather than a credit card surcharge. The two labels than gains, consumers are less likely to accept a surcharge than to forego a discount. As is to be expected, attempts to influence framing are common in the marketplace and in the political arena.

The evaluation of outcomes is susceptible to formulation effects because of the nonlinearity of the value function and the tendency of people to evaluate options in relation to the reference point that is suggested or implied by the statement of the problem. It is worthy of note that in other contexts people automatically transform equivalent messages into the same representation. Studies of language comprehension indicate that people quickly recode much of what they hear into an abstract representation that no longer distinguishes whether the idea was expressed in an active or in a passive form and no longer discriminates what was actually said from what was implied, presupposed, or implicated (Clark and Clark, 1977). Unfortunately, the mental machinery that performs these operations silently and effortlessly is not adequate to perform the task of recoding the two versions of the public health problem or the mortality-survival statistics into a common abstract form.

TRANSACTIONS AND TRADES

Our analysis of framing and of value can be extended to choices between multiattribute options, such as the acceptability of a transaction or a trade. We propose that, in order to evaluate a multiattribute option, a person sets up a mental account that specifies the advantages and the disadvantages associated with the option, relative to a multiattribute reference state. The overall value of an option is given by the balance of its advantages and its disadvantages in relation to the reference state. Thus, an option is acceptable if the value of its advantages exceeds the value of its disadvantages. This analysis assumes psychological—but not physical—separability of advantages and disadvantages. The model does not constrain the manner in which separate attributes are combined to form overall measures of advantage and of disadvantage, but it imposes on these measures assumptions of concavity and of loss aversion.

Our analysis of mental accounting owes a large debt to the stimulating work of Richard Thaler (1980, in press), who showed the relevance of this process to consumer behavior. The following problem, based on examples of Savage (1954) and Thaler (1980), introduces some of the rules that govern the construction of mental accounts and illustrates the extension of the concavity of value to the acceptability of transactions.

Problem 7: Imagine that you are about to purchase a jacket for $125 and a calculator for $15. The calculator salesman informs you that the calculator you wish to buy is on sale for $10 at the other branch of the store, located 20 minutes drive away. Would you make a trip to the other store?

This problem is concerned with the acceptability of an option that combines a disadvantage of inconvenience with a financial advantage that can be framed as a *minimal*, *topical*, or *comprehensive* account. The minimal account includes only the differences between the two options and disregards the features that they share. In the minimal account, the advantage associated with driving to the other store is framed as a gain of $5. A topical account relates the consequences of possible choices to a reference level that is determined by the context within which the decision arises. In the preceding problem, the relevant topic is the purchase of the calculator, and the benefit of the trip is therefore framed as a reduction of the price, from $15 to $10. Because the potential saving is associated only with the calculator, the price of the jacket is not included in the topical account. The price of the jacket, as well as other expenses, could well be included in a more comprehensive account in which the saving would be evaluated in relation to, say, monthly expenses.

The formulation of the preceding problem appears neutral with respect to the adoption of a minimal, topical, or comprehensive account. We suggest, however, that people will spontaneously frame decisions in terms of topical accounts that, in the context of decisionmaking, play a role analogous to that of ''good forms'' in perception and of basic-level categories in cognition. Topical organization, in conjunction with the concavity of value, entails that the willingness to travel to the other store for a saving of $5 on a calculator should be inversely related to the price of the calculator and should be independent of the price of the jacket. To test this prediction, we constructed another version of the problem in which the prices of the two items were interchanged. The price of the calculator was given as $125 in the first store and $120 in the other branch, and the price of the jacket was set at $15. As predicted, the proportions of respondents who said they would make the trip differed sharply in the two problems. The results showed that 68 percent of the respondents ($N = 88$) were willing to drive to the other branch to save $5 on a $15 calculator, but only 29 percent of 93 respondents were willing to make the same trip to save $5 on a $125 calculator. This finding supports the notion of topical organization of accounts, since the two versions are identical both in terms of a minimal and a comprehensive account.

The significance of topical accounts for consumer behavior is confirmed by the observation that the standard deviation of the prices that different stores in a city quote for the same product is roughly proportional to the average price of that product (Pratt et al., 1979). Since the dispersion of than gains, consumers are less likely to accept a surcharge than to forego a discount. As is to be expected, attempts to influence framing are common in the marketplace and in the political arena.

The evaluation of outcomes is susceptible to formulation effects because of the nonlinearity of the value function and the tendency of people to evaluate options in relation to the reference point that is suggested or implied by the statement of the problem. It is worthy of note that in other contexts people automatically transform equivalent messages into the same representation. Studies of language comprehension indicate that people quickly recode much of what they hear into an abstract representation that no longer distinguishes whether the idea was expressed in an active or in a passive form and no longer discriminates what was actually said from what was implied, presupposed, or implicated (Clark and Clark, 1977). Unfortunately, the mental machinery that performs these operations silently and effortlessly is not adequate to perform the task of recoding the two versions of the public health problem or the mortality-survival statistics into a common abstract form.

TRANSACTIONS AND TRADES

Our analysis of framing and of value can be extended to choices between multiattribute options, such as the acceptability of a transaction or a trade. We propose that, in order to evaluate a multiattribute option, a person sets up a mental account that specifies the advantages and the disadvantages associated with the option, relative to a multiattribute reference state. The overall value of an option is given by the balance of its advantages and its disadvantages in relation to the reference state. Thus, an option is acceptable if the value of its advantages exceeds the value of its disadvantages. This analysis assumes psychological—but not physical—separability of advantages and disadvantages. The model does not constrain the manner in which separate attributes are combined to form overall measures of advantage and of disadvantage, but it imposes on these measures assumptions of concavity and of loss aversion.

Our analysis of mental accounting owes a large debt to the stimulating work of Richard Thaler (1980, in press), who showed the relevance of this process to consumer behavior. The following problem, based on examples of Savage (1954) and Thaler (1980), introduces some of the rules that govern the construction of mental accounts and illustrates the extension of the concavity of value to the acceptability of transactions.

Problem 7: Imagine that you are about to purchase a jacket for $125 and a calculator for $15. The calculator salesman informs you that the calculator you wish to buy is on sale for $10 at the other branch of the store, located 20 minutes drive away. Would you make a trip to the other store?

This problem is concerned with the acceptability of an option that combines a disadvantage of inconvenience with a financial advantage that can be framed as a *minimal*, *topical*, or *comprehensive* account. The minimal account includes only the differences between the two options and disregards the features that they share. In the minimal account, the advantage associated with driving to the other store is framed as a gain of $5. A topical account relates the consequences of possible choices to a reference level that is determined by the context within which the decision arises. In the preceding problem, the relevant topic is the purchase of the calculator, and the benefit of the trip is therefore framed as a reduction of the price, from $15 to $10. Because the potential saving is associated only with the calculator, the price of the jacket is not included in the topical account. The price of the jacket, as well as other expenses, could well be included in a more comprehensive account in which the saving would be evaluated in relation to, say, monthly expenses.

The formulation of the preceding problem appears neutral with respect to the adoption of a minimal, topical, or comprehensive account. We suggest, however, that people will spontaneously frame decisions in terms of topical accounts that, in the context of decisionmaking, play a role analogous to that of "good forms" in perception and of basic-level categories in cognition. Topical organization, in conjunction with the concavity of value, entails that the willingness to travel to the other store for a saving of $5 on a calculator should be inversely related to the price of the calculator and should be independent of the price of the jacket. To test this prediction, we constructed another version of the problem in which the prices of the two items were interchanged. The price of the calculator was given as $125 in the first store and $120 in the other branch, and the price of the jacket was set at $15. As predicted, the proportions of respondents who said they would make the trip differed sharply in the two problems. The results showed that 68 percent of the respondents ($N = 88$) were willing to drive to the other branch to save $5 on a $15 calculator, but only 29 percent of 93 respondents were willing to make the same trip to save $5 on a $125 calculator. This finding supports the notion of topical organization of accounts, since the two versions are identical both in terms of a minimal and a comprehensive account.

The significance of topical accounts for consumer behavior is confirmed by the observation that the standard deviation of the prices that different stores in a city quote for the same product is roughly proportional to the average price of that product (Pratt et al., 1979). Since the dispersion of prices is surely controlled by shoppers' efforts to find the best buy, these results suggest that consumers hardly exert more effort to save $15 on a $150 purchase than to save $5 on a $50 purchase.

The topical organization of mental accounts leads people to evaluate gains and losses in relative rather than in absolute terms, resulting in large variations in the rate at which money is exchanged for other things, such as the number of phone calls made to find a good buy or the willingness to drive a long distance to get one. Most consumers will find it easier to buy a car stereo system or a Persian rug, respectively, in the context of buying a car or a house than separately. These observations, of course, run counter to the standard rational theory of consumer behavior, which assumes invariance and does not recognize the effects of mental accounting.

The following problems illustrate another example of mental accounting in which the posting of a cost to an account is controlled by topical organization:

Problem 8 ($N = 200$): Imagine that you have decided to see a play and paid the admission price of $10 per ticket. As you enter the theater, you discover that you have lost the ticket. The seat was not marked, and the ticket cannot be recovered.

Would you pay $10 for another ticket?
Yes (46%) No (54%)

Problem 9 ($N = 183$): Imagine that you have decided to see a play where admission is $10 per ticket. As you enter the theater, you discover that you have lost a $10 bill.

Would you still pay $10 for a ticket for the play?
Yes (88%) No (12%)

The difference between the responses to the two problems is intriguing. Why are so many people unwilling to spend $10 after having lost a ticket, if they would readily spend that sum after losing an equivalent amount of cash? We attribute the difference to the topical organization of mental accounts. Going to the theater is normally viewed as a transaction in which the cost of the ticket is exchanged for the experience of seeing the play. Buying a second ticket increases the cost of seeing the play to a level that many respondents apparently find unacceptable. In contrast, the loss of the cash is not posted to the account of the play, and it affects the purchase of a ticket only by making the individual feel slightly less affluent.

An interesting effect was observed when the two versions of the problem were presented to the same subjects. The willingness to replace a lost ticket increased significantly when that problem followed the lost-cash version. In contrast, the willingness to buy a ticket after losing cash was not affected by prior presentation of the other problem. The juxtaposition of the two problems apparently enabled the subjects to realize that it makes sense to think of the lost ticket as lost cash, but not vice versa.

The normative status of the effects of mental accounting is questionable. Unlike earlier examples, such as the public health problem, in which the two versions differed only in form, it can be argued that the alternative versions of the calculator and ticket problems differ also in substance. In particular, it may be more pleasurable to save $5 on a $15 purchase than on a larger purchase, and it may be more annoying to pay twice for the same ticket than to lose $10 in cash. Regret, frustration, and self-satisfaction can also be affected by framing (Kahneman and Tversky, 1982). If such secondary consequences are considered legitimate, then the observed preferences do not violate the criterion of invariance and cannot readily be ruled out as inconsistent or erroneous. On the other hand, secondary consequences may change upon reflection. The satisfaction of saving $5 on a $15 item can be marred if the consumer discovers that she would not have exerted the same effort to save $10 on a $200 purchase. We do not wish to recommend that any two decision problems that have the same primary consequences should be resolved in the same way. We propose, however, that systematic examination of alternative framings offers a useful reflective device that can help decisionmakers assess the values that should be attached to the primary and secondary consequences of their choices.

Losses and Costs

Many decision problems take the form of a choice between retaining the status quo and accepting an alternative to it, which is advantageous in some respects and disadvantageous in others. The analysis of value that was applied earlier to unidimensional risky prospects can be extended to this case by assuming that the status quo defines the reference level for all attributes. The advantages of alternative options will then be evaluated as gains and their disadvantages as losses. Because losses loom larger than gains, the decisionmaker will be biased in favor of retaining the status quo.

Thaler (1980) coined the term ''endowment effect'' to describe the reluctance of people to part from assets that belong to their endowment. When it is more painful to give up an asset than it is pleasurable to obtain it, buying prices will be significantly lower than selling prices. That is, the highest price that an individual will pay to acquire an asset will be smaller than the minimal compensation that would induce the same individual to give up that asset, once acquired. Thaler discussed some examples of the endowment effect in the behavior of consumers and entrepreneurs. Several studies have reported substantial discrepancies between buying and selling prices in both hypothetical and real transactions (Gregory, 1983; Hammack and Brown, 1974; Knetsch and Sinden, in press). These results have been presented as challenges to standard economic theory, in which buying and selling prices coincide except for transaction costs and effects of wealth. We also observed reluctance to trade in a study of choices between hypothetical jobs that differed in weekly salary (S) and in the temperature (T) of the workplace. Our respondents were asked to imagine that they held a particular position (S_1, T_1) and were offered the option of moving to a different position (S_2, T_2), which was better in one respect and worse in another. We found that most subjects who were assigned to (S_1, T_1) did not wish to move to (S_2, T_2), and that most subjects who were assigned to the latter position did not wish to move to the former. Evidently, the same difference in pay or in working conditions looms larger as a disadvantage than as an advantage.

In general, loss aversion favors stability over change. Imagine two hedonically identical twins who find two alternative environments equally attractive. Imagine further that by force of circumstance the twins are separated and placed in the two environments. As soon as they adopt their new states as reference points and evaluate the advantages and disadvantages of each other's environments accordingly, the twins will no longer be indifferent between the two states, and both will prefer to stay where they happen to be. Thus, the instability of preferences produces a preference for stability. In addition to favoring stability over change, the combination of adaptation and loss aversion provides limited protection against regret and envy by reducing the attractiveness of foregone alternatives and of others' endowments.

Loss aversion and the consequent endowment effect are unlikely to play a significant role in routine economic exchanges. The owner of a store, for example, does not experience money paid to suppliers as losses and money received from customers as gains. Instead, the merchant adds costs and revenues over some period of time and evaluates only the balance. Matching debits and credits are effectively cancelled prior to evaluation. Payments made by consumers are also not evaluated as losses but as alternative purchases. In accord with standard economic analysis, money is naturally viewed as a proxy for the goods and services that it could buy. This mode of evaluation is made explicit when an individual has in mind a particular alternative, such as "I can either buy a new camera or a new tent." In this analysis, a person will buy a camera if its subjective value exceeds the value of retaining the money it would cost.

There are cases in which a disadvantage can be framed either as a cost or as a loss. In particular, the purchase of insurance can also be framed as a choice between a sure loss and the risk of a greater loss. In such cases the cost-loss discrepancy can lead to failures of invariance. Consider, for example, the choice between a sure loss of $50 and a 25 percent chance to lose $200. Slovic et al. (1982) reported that 80 percent of their subjects expressed a risk-seeking preference for the gamble over the sure loss. However, only 35 percent of subjects refused to pay $50 for insurance against a 25 percent risk of losing $200. Similar results were also reported by Schoemaker and Kunreuther (1979) and by Hershey and Schoemaker (1980). We suggest that the same amount of money that was framed as an uncompensated loss in the first problem was framed as the cost of protection in the second. The modal preference was reversed in the two problems because losses are more aversive than costs.

We have observed a similar effect in the positive domain, as illustrated by the following pair of problems:

Problem 10: Would you accept a gamble that offers a 10% chance to win $95 and a 90% chance to lose $5?

Problem 11: Would you pay $5 to participate in a lottery that offers a 10% chance to win $100 and a 90% chance to win nothing?

A total of 132 undergraduates answered the two questions, which were separated by a short filler problem. The order of the questions was reversed for half the respondents. Although it is easily confirmed that the two problems offer objectively identical options, 55 of the respondents expressed different preferences in the two versions. Among them, 42 rejected the gamble in Problem 10 but accepted the equivalent lottery in Problem 11. The effectiveness of this seemingly inconsequential manipulation illustrates both the cost-loss discrepancy and the power of framing. Thinking of the $5 as a payment makes the venture more acceptable than thinking of the same amount as a loss.

The preceding analysis implies that an individual's subjective state can be improved by framing negative outcomes as costs rather than as losses. The possibility of such psychological manipulations may explain a paradoxical form of behavior that could be labeled the *dead-loss effect*. Thaler (1980) discussed the example of a man who develops tennis elbow soon after paying the membership fee in a tennis club and continues to play in agony to avoid wasting his investment. Assuming that the individual would not play if he had not paid the membership fee, the question arises: How can playing in agony improve the individual's lot? Playing in pain, we suggest, maintains the evaluation of the membership fee as a cost. If the individual were to stop playing, he would be forced to recognize the fee

CONCLUDING REMARKS

The concepts of utility and value are commonly used in two distinct senses: (a) *experience value*, the degree of pleasure or pain, satisfaction or anguish in the actual experience of an outcome; and (b) *decision value*, the contribution of an anticipated outcome to the overall attractiveness or aversiveness of an option in a choice. The distinction is rarely explicit in decision theory because it is tacitly assumed that decision values and experience values coincide. This assumption is part of the conception of an idealized decisionmaker who is able to predict future experiences with perfect accuracy and evaluate options accordingly. For ordinary decisionmakers, however, the correspondence of decision values between experience values is far from perfect (March, 1978). Some factors that affect experience are not easily anticipated, and some factors that affect decisions do not have a comparable impact on the experience of outcomes.

In contrast to the large amount of research on decisionmaking, there has been relatively little systematic exploration of the psychophysics that relate hedonic experience to objective states. The most basic problem of hedonic psychophysics is the determination of the level of adaptation or aspiration that separates positive from negative outcomes. The hedonic reference point is largely determined by the objective status quo, but it is also affected by expectations and social comparisons. An objective improvement can be experienced as a loss, for example, when an employee receives a smaller raise than everyone else in the office. The experience of pleasure or pain associated with a change of state is also critically dependent on the dynamics of hedonic adaptation. Brickman and Campbell's (1971) concept of the hedonic treadmill suggests the radical hypothesis that rapid adaptation will cause the effects of any objective improvement to be short-lived. The complexity and subtlety of hedonic experience make it difficult for the decisionmaker to anticipate the actual experience that outcomes will produce. Many a person who ordered a meal when ravenously hungry has admitted to a big mistake when the fifth course arrived on the table. The common mismatch of decision values and experience values introduces an additional element of uncertainty in many decision problems.

The prevalence of framing effects and violations of invariance further complicates the relation between decision values and experience values. The framing of outcomes often induces decision values that have no counterpart in actual experience. For example, the framing of outcomes of therapies for lung cancer in terms of mortality or survival is unlikely to affect experience, although it can have a pronounced influence on choice. In other cases, however, the framing of decisions affects not only decision but experience as well. For example, the framing of an expenditure as an uncompensated loss or as the price of insurance can probably influence the experience of that outcome. In such cases, the evaluation of outcomes in the context of decisions not only anticipates experience but also molds it.

REFERENCES

Allais, M., and Hagen, O., eds.

1979 *Expected Utility Hypotheses and the Allais Paradox.* Hingham, Mass.: D. Reidel Publishing.

Bernoulli, D.

1954 Exposition of a new theory on the measurement of risk. *Econometrica* 22:23–36. (Original work published in 1738.)

Brickman, P., and Campbell, D.T.

1971 Hedonic relativism and planning the good society. In M.H. Appley, ed., *Adaptation-level Theory: A Symposium.* Pp. 287–302. New York: Academic Press.

Clark, H.H., and Clark, E.V.

1977 *Psychology and Language.* New York: Harcourt Brace Jovanovich.

Erakar, S.E., and Sox, H.C.

1981 Assessment of patients' preferences for therapeutic outcomes. *Medical Decision Making* 1:29–39.

Fischhoff, B.

1983 Predicting frames. *Journal of Experimental Psychology: Learning, Memory and Cognition* 9:103–116.

Fischhoff, B., Slovic, P., and Lichtenstein, S.

1980 Knowing what you want: measuring labile values. In T. Wallsten, ed., *Cognitive Processes in Choice and Decision Behavior.* Pp. 117–141. Hillsdale, N.J.: Erlbaum.

Fishburn, P.C., and Kochenberger, G.A.

1979 Two-piece von Neumann-Morgenstern utility functions. *Decision Sciences* 10:503–518.

Gregory, R.

1983 Measures of Consumer's Surplus: Reasons for the Disparity in Observed Values. Unpublished manuscript, Keene State College, Keene, N.H.

Hammack, J., and Brown, G.M., Jr.

1974 *Waterfowl and Wetlands: Toward Bioeconomic Analysis.* Baltimore: Johns Hopkins University Press.

Hershey, J.C., and Schoemaker, P.J.H.

1980 Risk taking and problem context in the domain of losses: an expected-utility analysis. *Journal of Risk and Insurance* 47:111–132.

Kahneman, D., and Tversky, A.

1979 Prospect theory: an analysis of decision under risk. *Econometrica* 47:263–291.

1982 The simulation heuristic. In D. Kahneman, P. Slovic, and A. Tversky, eds., *Judgment Under Uncertainty: Heuristics and Biases.* Pp. 201–208. New York: Cambridge University Press.

Knetsch, J., and Sinden, J.

In press Willingness to pay and compensation demanded: Experimental evidence of an unexpected disparity in measures of value. *Quarterly Journal of Economics.*

March, J.G.

1978 Bounded rationality, ambiguity, and the engineering of choice. *Bell Journal of Economics* 9:587–608.

McNeil, B., Pauker, S., Sox, H., Jr., and Tversky, A.

1982 On the elicitation of preferences for alternative therapies. *New England Journal of Medicine* 306:1259–1262.

Payne, J.W., Laughhunn, D.J., and Crum, R.

1980 Translation of gambles and aspiration level effects in risky choice behavior. *Management Science* 26:1039–1060.

Pratt, J.W., Wise, D., and Zeckhauser, R.

1979 Price differences in almost competitive markets. *Quarterly Journal of Economics* 93:189–211.

Savage, L.J.

1954 *The Foundation of Statistics*. New York: Wiley.

Schlaifer, R.

1959 *Probability and Statistics for Business Decisions*. New York: McGraw-Hill.

Schoemaker, P.J.H., and Kunreuther, H.C.

1979 An experimental study of insurance decisions. *Journal of Risk and Insurance* 46:603–618.

Slovic, P., Fischhoff, B., and Lichtenstein, S.

1982 Response mode, framing, and information-processing effects in risk assessment. In R. Hogarth, ed., *New Directions for Methodology of Social and Behavioral Science: Question Framing and Response Consistency*. Pp. 21–36. San Francisco: Jossey-Bass.

Thaler, R.

1980 Toward a positive theory of consumer choice. *Journal of Economic Behavior and Organization* 1:39–60.

In press Using mental accounting in a theory of consumer behavior. *Journal of Marketing*.

Tversky, A.

1977 On the elicitation of preferences: descriptive and prescriptive considerations. In D. Bell, R.L. Kenney, and H. Raiffa, eds., *Conflicting Objectives in Decisions. International Series on Applied Systems Analysis*. Pp. 209–222. New York: Wiley.

Tversky, A., and Kahneman, D.

1981 The framing of decisions and the psychology of choice. *Science* 211:453–458.

von Neumann, J., and Morgenstern, O.

1947 *Theory of Games and Economic Behavior*. 2nd ed. Princeton: Princeton University Press.

Discovering the Mind at Work

Changing Views of Cognitive Competence in the Young

ROCHEL GELMAN and ANN L. BROWN

It was once commonly thought that the newborn child cannot hear, see, or smell; that the first year of life is spent in a blooming-buzzing confusion; and that infants lack ability to form complex ideas about the world. For much of this century, most experimental and developmental psychologists accepted the traditional thesis that the newborn's mind is a blank slate (or *tabula rasa*) upon which the record of experience is gradually impressed. It was further held that language is an obvious prerequisite for any abstract thought (e.g., Vygotsky, 1962; Whorf, 1956), so in its absence, a baby could not have knowledge of anything other than sensations. Since babies are born with an extremely limited repertoire of behavior and spend most of their early months asleep, they certainly appear passive and unknowing; there is no obvious way for them to demonstrate otherwise.

But challenges to this view arose. On the theoretical side, it is hard to overestimate the impact of Piaget, Chomsky, Simon, and the Gibsons, who profoundly influenced psychologists' ideas of what to look for and how to characterize the child. These new theoretical views stimulated innovative research programs with the very young. It became clear that with carefully

Partial support for the preparation of this chapter came from fellowships to the authors while at the Center for Advanced Study in the Behavioral Sciences, Stanford, California. While there both authors received funds from the Spencer Foundation, and Rochel Gelman was also supported by the Alfred P. Sloan Foundation and a National Institute of Child Health and Development (NICHD) Fellowship F32 HD-06623. Additional support for the work came from National Science Foundation grant BNS-8140573 to Gelman and NICHD grants HD-06864 and HD-05951 to Ann Brown.

designed methods one could find ways to pose rather complex questions to infants and young children. A substantial new body of data has now accumulated about the capacities of infants and young children and stands in contrast to the older emphasis on what they lack. From these data a contemporary view has emerged that the very young can be competent, active agents of their own conceptual development. In short, the mind of the young child has come to life.

This essay is divided into four parts. First, we introduce the seminal theoretical ideas that have influenced psychologists' conceptions of the child's emergent mind. Next, we delineate some of the evidence in support of infant and preschool cognitive competence and illustrate some of the methods developed to make the study of young minds plausible. Finally, we ask how this putative youthful brilliance interacts with formal learning tasks in school, emerging with a seeming paradox: Young children seem to know more than we thought possible, but older children in schools seem to be much less competent than was once assumed. The natural learning settings of young children are contrasted with the formal environments they encounter at school, and we see that instructional programs that capitalize on young children's natural propensities to create and test theories can significantly accelerate learning.

THEORETICAL BACKGROUND

The first step away from the empiricists' ''tabula rasa'' view of the infant mind was taken by the Swiss psychologist Jean Piaget. Beginning in the 1920s, Piaget argued for the need to postulate complex cognitive structures in the young human mind, which empiricist accounts of human thought had tended to play down or deny. Piaget did not think that human infants are born with innate cognitive structures, but rather that structures develop due to the child's ever-present tendency to engage the environment actively, interpreting it in accordance with progressively changing cognitive ''schemes.'' From close observations of infants and careful questioning of children, he concluded that cognitive development proceeds through certain stages, each involving radically different cognitive schemes, so that sometimes young children even form practical convictions contrary to those held by older children and adults.

While Piaget observed that infants actually seek environmental stimulation that promotes their intellectual development, he thought that their initial representations of objects, space, time, cause, and self are constructed only gradually during the first two years. He concluded that the world of young infants is an egocentric fusion of the internal and external worlds, and that the development of an accurate representation of physical reality depends on the gradual coordination of schemes of looking, listening, and touching. Piaget thought that for many months the infant does not realize that an object producing a given sound is the same as an object that looks a certain way. The very young infant, up to 10 months or so, was said to think that an object exists only as long as she can touch, hear, or see it; once out of direct sensory contact, it ceased to exist. From this view, it followed that babies do not represent an independent space in which three-dimensional objects exist. In this regard, Piaget's account of infant cognition is actually close to being empiricist; still the position that cognitive schemes are actively constructed rather than passively impressed separates him from empiricists (or in modern terminology, behaviorists).

Noam Chomsky (1957), focusing on language, proposed that the human mind is innately prepared to learn language without needing much help from the environment. He provided explicit hypotheses about the nature of the language structures that produce and comprehend language, an account that held out the promise of explaining how young children can say things they have never heard, e.g., "I'm unthirsty," "I have two footses," "I wented home." Chomsky's hypotheses are still controversial (Wanner and Gleitman, 1982), but the effect of his work gave strong impetus to a "nativist" account of mental abilities, which maintains that humans are born with conceptual structures that guide the acquisition of knowledge about the world.

Like Piaget, the Gibsons have maintained that infants actively explore the environment, but in sharp contrast, they deny that the infant slowly constructs the world. They maintain that, shortly after birth, the infant's world is a remarkably veridical one, filled with three-dimensional objects in real space, not unconnected elementary sensations. They support their view with findings that neonates integrate sight and sound and respond as if they assume that the world is out there waiting to be explored. The Gibsons assign a role to learning but propose that it proceeds rapidly due to the initial availability of exploration patterns that can yield accurate information about objects and events.

Simon (1972) and his colleagues (e.g., Klahr and Wallace, 1976) helped introduce a somewhat different perspective, that development means overcoming information-processing constraints, such as limited short-term memory capacity and lack of general knowledge. Those working in the information-processing tradition focused both on the possibility that early failures in completing Piagetian tasks are due in part to limits on processing capacity and the conditions under which children actively employ strategies for problem solving and knowledge acquisition.

All these theoretical developments challenged the empiricist account and influenced the direction of research in developmental psychology. The claim that young children have different mental structures and ideas about the world was taken up in investigations of their concepts, strategies, and problem-solving abilities. These studies led to the conclusion that, despite the many differences between young and old, the young have remarkable abilities to participate actively in their acquisition of knowledge.

STUDYING INFANT KNOWLEDGE

Because infants are so limited physically, experiments to find out what they know and how they think have had to find methods suitable to the level of infant motor capabilities. A good example is a method used by Kalnins and Bruner (1973). They showed 5- to 12-week-old infants a silent color film and gave the infants a pacifier to suck, the nipple of which was connected to a pressure switch controlling the projector lens. The infants quickly learned to suck at a given rate to bring the movie into focus, showing not only that they were capable of and interested in learning how to control their own sensory environment but also that they preferred a clear image to a blurry one.

A second method demonstrates—and depends on—an infant's thirst for novelty. The "habituation paradigm" involves presenting babies with a stimulus—a picture, sound, or series of sounds—to which the baby attends either by looking at it, turning to it, or doing something to keep the stimulus on. Over a series of trials, infants, like everyone else, stop responding to repeated presentations of the same stimulus; that is, they *habituate*. They recover interest if a recognizably different stimulus is presented. For example, four-month-old infants will suck vigorously when first introduced to the phoneme (speech sound) "ba," then gradually lose interest and stop sucking in response to it. But when presented a different phoneme, "pa," they resume sucking (Eimas et al., 1971).

Fantz (1961, 1966) directed attention to the power of the preference method to study infants' tendency to explore. He determined what infants looked at by watching their eyes closely. Infants lying on their backs in his laboratory could look up to the left or right at, for example, a bull's eye and a checkerboard. The experimenter recorded whether and for how long the baby looked left or right. Even newborns chose to look at patterned displays over homogenous gray ones. Infants generally prefer somewhat novel displays over ones they have seen before (e.g., Kagan et al., 1978; Kessen et al., 1972).

Studies like these do more than simply show that infants actively select experiences; they can also tell us what the infant is capable of perceiving and knowing. Recovery of interest in a novel speech sound could not occur if infants could not recognize the rather subtle difference between "pa" and "ba." (See Aslin et al., 1983.) The same holds for visual preferences. Discovering that very young infants can see, hear, smell, and be particular about what exactly they perceive led to an emboldened attitude about the kinds of experimental questions that could be asked. The answers about infant understanding of the physical and numerical properties of objects have been quite remarkable.

Early Knowledge of Objects

Piaget concluded that before infants could know about objects, they would have to discover regularities between their sensations and actions, then gradually integrate the sense-action schemes formed when they touched, heard, and looked at objects, and finally come to appreciate the object as a separate reality in the external world. Like the empiricists, Piaget thought infants responded to the immediate stimuli, i.e., flashes of light on the retina or sound waves in the eardrum, long before they recognized sources of stimuli.

Recent experiments (Gibson and Spelke, 1983; Harris, 1983) have told a different story. For example, Spelke (1976) used visual-preference methods to determine that four-month-old infants already integrate the sight and sound of an event. Infants were shown two films projected side by side—a person playing peek-a-boo and a hand beating a tambourine. The sound accompaniment of one film was fed to a hidden loudspeaker placed midway between the films. The babies reliably preferred to look at the movie corresponding to the sound source. Other research indicates that babies are born with a tendency to turn to a sound and visually search for something there (Field et al., 1980; Mendelson and Haith, 1976; Wertheimer, 1961).

This integrative capacity extends beyond auditory and visual properties to include the sense of touch. In recent experiments, Gibson and Walker (1984) gave one-month-old infants either a hard lucite cylinder or a lookalike soft sponge cylinder to explore with their mouths. The experimenter then showed each infant both cylinders, squeezing the spongy cylinder in one hand and rotating the hard cylinder with the other hand. The infants preferred to look at whichever cylinder had not previously been explored orally, showing a capacity to integrate what they saw with what they had mouthed. Meltzoff and Borton (1979) reported similar findings with objects that were smooth or had tiny nobs on their surface.

These findings establish two important points about the cognitive structure that infants employ to interpret sensory input from objects: (a) They endow objects with properties, such as rigidity, that transcend sensory modality; and (b) infants can appreciate such properties even when they are not acting on the objects.

Von Hofsten (1980) provides further evidence that babies know things about objects before they can successfully act on them. Around four months of age, infants are able to reach out and grasp objects. At the same time, without having had the experience of successfully catching a moving object, they also anticipate trajectories correctly and move their hand toward the spot where a moving object will be. This would appear to require reckoning of the velocity and direction of the object, foreknowledge of the time that the arm movement will take, and ability to combine these in calculating an intercept.

Piaget noted the considerable difficulty infants have with occluded objects. When infants four to eight months old are shown an interesting object, they reach for and grasp it and even follow its fall to the floor. But they stop reaching or looking if the object disappears behind a barrier. Infants 8 to 12 months old will seek and retrieve an object they see someone cover, but they show an odd tendency referred to as the "A not B error." If the baby sees an object taken from behind one barrier, A, and, *while the baby watches*, moved behind another barrier, B, the baby searches only behind the first barrier (A)! Piaget concluded that "the object is still not the same to the child as it is to us: a substantial body, individualized and displaced in space without depending on the action context in which it was inserted" (Piaget, 1954:64). Recent research suggests that the A not B error may be confined to particular experimental situations (Bjork and Cummings, 1984; Sophian, 1984). After all, if babies can match what they mouth with what they see, distinguishing between solid and spongy substances, they must be sensitive to objects as substantial bodies.

As further evidence for this view, Baillargeon et al. (in press) showed five-month-old infants a screen that rotated toward and away from them through an arc of 180 degrees. Once the babies were habituated to the rotating screen, a yellow cube was placed alongside the screen for two trials of viewing. Then the cube was placed behind the screen; on alternating trials, the infant saw either a screen that once again rotated through the full 180-degree arc (and at least from the adult perspective seemed to crush the covered object) or a screen that rotated through only a 120-degree arc, stopping at the angle at which its further rotation would be blocked by the presence of a solid object behind the screen. (One-way mirrors and varied lighting accomplished the visual effects.) Although the infants had previously habituated to the full rather than partial rotation, they nonetheless looked longer at the full rotation, treating the habituated event as even more novel than one they had never seen before. These results suggest the babies expect solid objects to persist even when no longer in sight.

In short, considerable research with young infants has shown that they treat objects and events as sources for multiple kinds of sensory input, and that they recognize in objects properties such as rigidity and solidity that transcend specific sensory modalities.

Abstract Concepts

Many theories of cognition have assumed that language is necessary to abstract properties common to a set of objects. Premack (1976) tellingly refuted this thesis when he showed that once a chimpanzee had learned the symbol for apple, it could apply that symbol to various parts of an apple (seed, peel, etc.). Preverbal human infants also recognize properties common to sets of nonidentical objects. Ross (1980), for example, habituated one- and two-year-olds to one of five classes of items: O shapes, M shapes, furniture, men, and food. Then children were shown another item from the same class or an item from a novel class. They preferred the item from the novel class. The children's ability to recognize category membership was uncorrelated with their ability to supply a verbal label for the category.

Number is a property of sets divorced from any description of the objects themselves. Hence, it is often treated as the ultimate in abstraction. A variety of results indicate that infants abstract number from visual displays of two, three, and sometimes four items (Starkey and Cooper, 1980; Starkey et al., in press; Strauss and Curtis, 1981). For example, six- to nine-month-old infants became habituated to color photographs of either two or three assorted common household items, e.g., sponge, cloth, vase, comb, apple, etc. (each trial displayed different items). Infants who were habituated to two-object displays then looked longer at three-item ones, and vice versa. Infants even abstract number intermodally (Starkey et al., 1983). They prefer to look at the one of two displays that matches the number of drumbeats (two or three) they hear emanating from a centrally placed loudspeaker.

Summary

We have sampled the evidence that infants are not passive, unstructured receivers of environmental input. Soon after birth they reveal an impressive degree of implicit conceptual structure allied to active learning endeavors. They behave as if they recognize that objects are independent of themselves, having size and solidity, and are specified intermodally. They reveal sensitivity to some properties of moving objects and form concepts about some abstract properties of sets. It is not at all obvious why infants bother to attend to the number of items they see or hear. But it looks as if human infants come prepared to learn quickly about objects and certain concepts, including number. These early competences provide a base from which much natural learning proceeds during the preschool years. Acquisition of knowledge in these natural domains is guided as much by the availability of implicit structures and principles that guide the child's active learning about the nature of objects and events, causes, number, etc. as by the availability of a supportive environment.

PRESCHOOL THOUGHT

Principles About Numbers, Causes, and Objects

Preschool thought and its development are much influenced by implicit knowledge of fundamental principles governing the determination and manipulation of numbers, the character of physical causality, and the differences between animate and inanimate objects.

Number Many preschoolers spontaneously count collections soon after they learn to talk. Gelman and Gallistel (1978) propose that even these very young children make implicit use of some, if not all, of five principles of counting: (1) The tags used in counting must be placed in one-to-one correspondence with the items counted; (2) the tags must be drawn in order from a stably ordered list; (3) the last tag used represents the number in the set (cardinality); (4) the order in which the items are tagged is irrelevant; and (5) sets of arbitrary composition may be counted. What evidence is there for this view?

First, counting behaviors in young children are systematic, even when they use nonstandard tags or orderings. For example, Gelman and Gallistel (1978) report a two-and-one-half year old who said "one, two" when counting a two-item array and "one, two, six" when counting a three-item array (the one-one principle). The same child used her own list over and over again (stable order principle) and repeated her last tag when asked how many items she had (the cardinality principle). Such nonstandard lists in counting are like the systematic errors made by young language learners (e.g., "I runned"); just as the occurrence of such language errors implies use of language rules by the very young, so the occurrence of stable nonstandard lists can be taken as evidence of implicit counting principles. Further evidence for implicit counting principles is found in the fact that young children spontaneously self-correct their own and others' counting errors (Gelman and Meck, 1983) and often are inclined to count without any request to do so. Such behaviors point to a representation that monitors and motivates performance (Greeno et al., 1984).

Other studies have shown that preschool children solve simple arithmetic problems by using counting strategies they invent (Groen and Resnick, 1977; Siegler and Robinson, 1982). To illustrate, Groen and Resnick (1977) taught four- and five-year-old children to solve addition problems of the form x + y = ? by counting out x blocks, counting out y more blocks and then counting the combined set. Children who practiced their addition over several weeks got better. More surprising is that over half of them invented a better way of solving the problems; counting on from whichever was the larger of the two values in the problem. To account for such inventions, it is necessary to postulate the use of something like an implicit principle of commutativity.

Finally, the preschool child also understands that addition and subtraction, unlike displacement, rearrangement, or item substitution, alter numerosity. This has been shown in "magic" experiments where a child is confronted with unexpected alterations in the sets used in a kind of shell game (Gelman, 1977). In these experiments children between the ages of three and five first learn to find plates holding different numbers of objects, e.g., two and three, underneath each of two cans. Then they discover surreptitious changes in the number, type, or arrangement of items in one array. Those children who encountered irrelevant changes deemed them such and those who encountered the effects of relevant transformations pronounced them relevant. For example, changes in number elicited considerable surprise, e.g., "Eeeeee, how did that happen?" Further, the children postulated the relevant transformation, e.g., "One gone—Jesus Christ came and took it." They also could indicate what number they expected, what number they actually encountered, and what arithmetic operation would have to be performed to "fix" the game—in this case, addition.

Hence we see early implicit understanding of number, addition, and subtraction. We will later ask why this competence does not guarantee easy learning of mathematics in school.

Causality The suggestion that young children work with implicit notions of cause will surprise those familiar with Piaget's work on the development of the child's conception of causality. In one set of inquiries, Piaget asked children to explain a variety of natural and mechanical phenomena, e.g., the cycle of the moon, floating objects, the movement of clouds, the operation of steam engines, and bicycles, etc. Analysis of the explanations led Piaget to characterize the young child's thought as fundamentally *precausal.* He wrote, "Immediacy of relations and absence of intermediaries . . . are the two outstanding features of causality around the age of four-five" (Piaget, 1980:268).

Piaget's conclusion that a concern for mechanism is completely lacking in the preschooler is contradicted by several later lines of experimental research. For example, Shultz (1982) showed two-year-olds the cause-effect sequence of turning on a blower that then blew out a candle. He then showed them two blowers, each surrounded on three sides by a plexiglass shield, the critical difference being whether the open side was facing the candle. If considerations of mechanism did not influence young children, they would choose randomly between the two blowers as potential causes of extinguishing the candle. Instead they systematically chose the unblocked blower. Similar findings were reported for the transmission of sound from a tuning fork or light from a battery. Preschoolers consistently took note of barriers that would stop the transmission of prerequisite energy. Comparable results were obtained by Shultz with schooled and unschooled Mali children in West Africa.

Other lines of evidence support the conclusion that preschoolers on many occasions do reveal an implicit concern for cause. Hood and Bloom (1979) note a ubiquitous tendency for children to seek causal accounts of what happens and how things work. Bullock (in press) showed that young children distinguish plausible from implausible mechanisms. At the start of her experiment, a rolling steel ball and a rolling light (produced as is the moving light effect in a movie marquee) moved simultaneously down parallel runways and disappeared together at the same time into an adjoining box. After a brief delay Snoopy jumped out of the box. Children were asked to identify the cause of Snoopy's jumping. They reliably named the ball. Since both preceding events were coterminous and redundant, the children should have shown no preference for one event over the other if they considered only temporal and spatial contiguity when reasoning about causality. Their preference for the ball (an object with momentum and kinetic energy) can be taken as evidence that they were concerned with plausibility of mechanism.

Findings like these have led many (e.g., Bullock et al., 1982; Koslowski et al., 1981; Shultz, 1982) to a view that preschool children work with a set of implicit assumptions about physical causality, including the crucial one that mechanisms mediate cause and effect relations. Guided by these implicit assumptions, they learn rapidly about their world; but, as we shall see, this does not guarantee the acquisition of scientifically correct theories.

Objects An early concern for mechanism may explain why preschool children are able to separate animate and inanimate objects (Carey, 1985a; Gelman et al., 1983; Keil, 1979). For example, three- to five-year-olds, asked whether a rock, a doll, and a person could walk, typically answered that a rock cannot walk because it has no feet; that a doll cannot walk unless someone pushes it, because its feet are only pretend; and that people can walk by themselves. In other words, inanimate objects cannot cause themselves to move, whereas animate ones can. Even infants treat animate and inanimate objects as belonging to different categories; they become very upset when a human stands still and fails to respond to them (e.g., Field, 1978; Tronick et al., 1975) but do not do so in the presence of inanimate objects. These findings have led theorists to postulate that humans are disposed to treat animate and inanimate objects separately at birth. Since infants respond differently to moving malleable objects than moving solid objects, this may reflect early recognition of fundamental differences in the way animate and inanimate objects move.

Making Plans and Strategies

This review of preschoolers' knowledge of numbers, causes, and objects only scratches the surface of evidence that the young are more competent than we once presumed. For example, there is compelling evidence that preschoolers' interest in and recall of stories reflects the availability of story grammars (Mandler, 1983; Stein and Trabasso, 1982); that preschoolers can systematically classify (Rosch et al., 1976); that they can be logical (Braine and Rumain, 1983); that they represent knowledge with a variety of coherent structures (Keil, 1981; Markman, 1981; Nelson and Gruendel, 1981); that children this age can take account of the perspective of an observer other than themselves (Lempers et al., 1977; Shatz, 1978); and even that congenitally blind children have Euclidean representations of space (Landau et al., 1981).

Given all this evidence, it should not be surprising to discover that preschool children are often strategic and planful when acquiring knowledge structures. This was not always assumed, however. The dominant developmental theories of the 1960s argued that a major shift in the quality of children's learning occurred between the ages of five and seven years; prior to that shift, children's learning was seen as primarily nonstrategic, passive, and context dependent; only after the shift was children's learning thought to become increasingly strategic, active, and flexible.

These ideas were not advanced in the abstract. An enormous empirical base backed them up (Stevenson, 1970; White, 1965, 1970), but much of this data base was built using experimental designs that were not suitable for young children. Given cognitive exercises designed for school-age students, preschoolers typically performed abysmally, if at all, thus confirming theoretical claims of their incompetence.

Systematic attempts to find more suitable ways to test the competence of younger children began in the 1970s (for a discussion see Brown and Deloache, 1978; Donaldson, 1978; Gelman, 1978). We will present some selections from the growing evidence that preschool children behave strategically, often direct their own learning, and actively create, test, and refine their own theories of the world around them. Driving this reassessment of preschool learning has been a greater consideration of the context in which it is observed.

Strategies for Remembering A great deal of research in the 1960s and early 1970s was concerned with the development of school-age children's strategies for enhancing memory (Brown, 1975; Flavell, 1970a, 1970b; Kail and Hagen, 1977). A central theme was that preschool children would differ from grade-school children on tasks that demand a great deal of strategic ingenuity. Young children, failing to devise strategic plans, would be at a considerable disadvantage on tests of deliberate memory, whereas older children would display increasing competence, primarily because they deploy more and more effective learning strategies. But laboratory and school tests of deliberate memory do not translate readily into the contexts in which young children naturally practice their emergent retention skills. That four-year-olds tend to be at a loss when asked to reproduce lists of digits or letters does not mean that they completely lack the ability to plan for future memory demands.

How, for example, could one reconcile the diagnosis of nonstrategic learning with the following description of three-year-olds anticipating a memory test? Surreptitiously observing children as they attempted to remember which of several containers concealed a toy dog, Wellman et al. (1975) found clear evidence of rehearsal (looking at the target container and nodding yes, looking at the nontarget containers and nodding no), retrieval cueing (resting their hands on the correct container or moving it to a salient position), and focused attention (looking fixedly at the correct hiding place). The children refused to be distracted until they were permitted to retrieve the lost dog. These efforts were rewarded: children who prepared actively for retrieval did remember better.

DeLoache et al. (1985) found even earlier evidence of planning for future retrieval. Children 18 to 24 months old were observed playing a hide-and-seek game; an attractive toy (Big Bird) was hidden in a variety of locations in a laboratory waiting room, such as behind a pillow on a couch. A timer was set to indicate the retrieval interval of, for example, five minutes; when the bell rang, the child could retrieve the toy. Far from waiting passively, the children interrupted their play to engage in activities indicating they were still preoccupied with the memory task: talking about the toy, pointing to the hiding place, or attempting an illegal peek. The children did not engage in these "keep-alive" activities if the toy remained partially visible during the retention interval or if the experimenter was responsible for remembering the location. Many other examples of early strategic competence could be cited; precocious strategic competence is not limited to attempts at remembering.

Theory-building Typecasting the young child as a passive learner also leads to a view of the young as being dependent on others for opportunities to learn, most notably parents, peers, or teachers. Some have argued that much, if not most, cognitive growth is a result of children internalizing cognitive activities that they originally witness in others (Laboratory of Comparative Human Cognition, 1983; Rogoff and Wertsch, 1984).

Interesting and important though guided learning situations may be, it is clear that much of the time children are also actively involved in orchestrating their own learning. Children learn in situations where there is no obvious guidance, no feedback other than their own satisfaction, and no external pressure to improve or change. They act like scientists, creating theories-in-action (Karmiloff-Smith, 1985) that they challenge, extend, and modify quite on their own. The child is not only a problem-solver but a problem-creator, a metaphor much in keeping with scientific thinking.

Some of the best evidence of self-motivated learning comes from situations in which children are observed as they operate on a problem, over considerable periods of time, quite without external pressure, seemingly with no motivation other than to improve the theory on which they are working. Consider the behavior of 24- to 48-month-old children engaged in free play with a set of nested cups (DeLoache, Sugarman, and Brown, in press). Although the children saw the cups nested before they began to play, there was no real need to renest them; however, they did so, working long and hard in the process.

The most primitive activity, used frequently by children younger than 30 months, was *brute force*. When a large cup was placed on a smaller one, the children would repeatedly twist, bang, or press down hard on the nonfitting cup. A second approach used by some of the younger children was that of *local correction*. After placing two nonfitting cups together, the child separated them and tried to find a replacement for only one cup, a minimal restructuring involving the relation between only two cups at a time. A third characteristic ploy of children younger than 30 months was to *dismantle the entire set* and start again whenever a cup did not fit.

Older children (30 to 42 months) faced with a nonfitting cup engaged in strategies that involved *consideration of the entire set* of relations in the stack. For example, one sophisticated strategy was *insertion*; the children took apart the stack at a point that enabled them to insert a new cup correctly. A second strategy, *reversal*, was also shown by older children. After placing two nonfitting cups together, the child would immediately reverse the relation between them (5/4 immediately switched to 4/5).

The rapidly executed reversal strategy was not shown by the younger group. Some young children would repeatedly assemble, for example, cups 4–1, starting with 4 as a base and then inserting 3, 2, 1. Then they encountered the largest cup, that is, 5, and attempted to insert it on top of the completed partial stack, pressing and twisting repeatedly. When brute force failed, they would dismantle the whole stack and start again. Similarly, having assembled 1, 2, 4, and 5, and then encountering 3, the younger children's only recourse was to begin again.

The young learners progress from piecemeal activities and local fixup ploys to a thoughtful consideration of the relation among elements of the whole problem. There is evidence that this progression reflects a general learning mechanism in action that children of many ages use when faced with novel construction problems. A similar progression is seen in older (four to seven years) children attempting to construct a railway circuit (Karmiloff-Smith, 1979) and even in adolescents refining the processes of written composition (Scardamalia, 1984). It is also important to note that the development on any one task is not completely age-governed in that children left to work on the problem over short periods of time (hours, days, etc.) show the same developmental progression from immature to mature activities that characterizes the cross-age descriptions of initial attacks on the problem (Brown, Kane, and DeLoache, work in progress; Karmiloff-Smith and Inhelder, 1974–1975).

In most of the above examples, children left to work with the problem unaided create solutions, modify their own answers, correct their errors, and develop more mature strategies on their own. Perhaps more impressive cases are those in which children persist after an adequate solution has been reached. Reorganization and improvement in strategies is not solely a response to failure, but often occurs when the child seeks to improve quite adequate functioning procedures. In these cases, it is not failure that directs change but success that the child wishes to refine and extend.

Consider, for example, the group of four- to seven-year-olds who were asked to balance rectangular wooden blocks on a narrow metal rod (Karmiloff-Smith and Inhelder, 1974–1975). These were no ordinary blocks, however. Standard blocks had their weight evenly distributed, and could therefore be balanced at the geometric center. Weighted blocks had the weight of each "side" varied either conspicuously (by gluing a large square block to one end of the base rectangular block) or inconspicuously (by inserting a hidden weight into a cavity on one end); the geometric center rule would not work for these blocks.

At first, the children made the blocks balance by *brute trial and error*. This ploy was obviously successful; the children balanced each block in turn. This early errorless but unanalyzed phase was spontaneously supplanted by the emergence of strong theories-in-action directed at uncovering the rules governing balance in the miniature world of these particular blocks. Unfortunately, they were incomplete hypotheses that produced errors. A common early theory was to concentrate exclusively on the geometric center and attempt to balance all blocks in this fashion. This works only for standard blocks; the weighted blocks were discarded as exceptions (''impossible to balance''), even though the child had previously balanced them all.

After this theory was well established, the child became discomfited by the number and regularity of errors. A new juxtaposed theory was then developed for conspicuously weighted blocks. For these, the children compensated for the weight that was obviously added to one end and adjusted the point of balance accordingly. For a time, however, length and weight were considered independently; standard blocks were balanced by the geometric center rule and conspicuously weighted blocks by the rule of ''estimate weight first and then compensate.'' Hidden weight problems still generated errors; these blocks looked identical to the standard ones and were therefore subjected to the geometric center rule; when they did not conform, they were discarded as anomalies, ''impossible to balance.''

Now the young theorists were made uncomfortable by the remaining exceptions and began to seek a rule for them. In so doing, a reorganization was induced that resulted in a single rule for all blocks. The children paused before balancing any block and roughly assessed the point of balance. Verbal responses reflected their consideration of both length and weight, e.g., ''You have to be careful, sometimes it's just as heavy on each side and so the middle is right, and sometimes it's heavier on one side.'' *After* inferring the probable point of balance, and only then, did the child place the block on the bar.

For all of these examples we can ask, why do children bother? Implicit in the situation is the goal that the cups should be nested, the railway constructed, or the blocks balanced; but the children are free to abandon their efforts whenever they like. They persist, however, for long periods, even in the face of frustration and even when an adequate partial solution has been reached.

Pressure to work on adequate partial theories, to produce more encompassing theories, is very similar to what occurs in scientific reasoning. Like the scientist, it is essential that the child first develop simple theories that they perfect and control before they entertain more encompassing complex hypotheses. Karmiloff-Smith and Inhelder refer to this as creative simplification. By ignoring some of the complicating factors initially, the child can begin to construct theories that achieve partial success. Progress comes only when the inadequate partial theory is well established and the learner scientists are able to discover new properties that in turn make it possible for new theories to be constructed.

Summary

The studies reviewed in this section make it clear that a universal diagnosis of young children as passive learners, with little control of their own cognitive growth, does serious injustice to their ingenuity. Faced with problems to solve, where they are interested in the outcome and understand the goal, even two-year-olds behave like scientists, actively exploring the environment, testing theories in action, and modifying approaches to problems as a result of experience.

This is not to claim that two-year-olds possess problem-solving abilities comparable to those of the adult, or even of the eight-year-old. Nor is it to claim that preschool theory building is comparable to scientific reasoning perfected during the adolescent, college, and later years. Precursors of active, systematic problem solving emerge early in the child's life, but there are limits on the young child's theory building, and they can have considerable difficulty harnessing their natural proclivities in settings of formal education. These matters are taken up in the fourth section of this essay.

THE TRANSITION TO FORMAL SCHOOLING

Incomplete Knowledge

Young children develop and test theories about the nature of objects, numbers, causality, etc., but these theories are implicit, partial, limited, and sometimes wrong. Their further development depends to some extent on the kind of structured input offered in school, input that makes these theories more precise and explicit.

For example, young children sense that animate and inanimate objects differ, but a great deal more knowledge is needed to develop organized biological theories. Preschool children do not think of animals in the same way that older children and adults in this culture do, i.e., as sharing certain defining biological characteristics (Carey, 1985b). For example, if preschool children are taught that a person has a stomach, they allow that other animals do as well, but inanimate objects do not. However, if they are instead taught that a dog has a stomach, they do not necessarily attribute this to people and other animals. Carey postulates that young children's theory of animates is based on their theory of people and only later on a biological theory.

In a similar vein, preschoolers' understanding of number extends to ›me but not all situations. For preschoolers, a number is what you get / counting objects in a set. This is a good theory to a degree, but it ıs limits; for example, it seems to hinder the expansion of the children's ımber system to include zero and negative numbers. Zero is not a tag at one applies to an item in a set being counted; nor of course is minus ıe (Evans, 1983).

Misconceptions about the centrality of counting must contribute to systmatic errors that elementary school children make in subtraction problems: ey are strongly inclined to subtract the smaller digit from the larger, and ırrying across zero presents children with unusual difficulty (Brown and anLehn, 1982; Lindvall and Ibarra, 1981).

Another characteristic difference between the theories of the young and der individuals is in their explicitness. The theories entertained by prehoolers are almost always implicit; they cannot be articulated but nevereless seem to determine beliefs and actions in a given domain. To illustrate e power of an implicit theory: All English speakers have extensive implicit ıowledge of English syntax, knowledge that constrains what we say and ıderstand, but that, in the absence of linguistic instruction, we cannot ticulate. No one ever says: "Who did John see Mary and?" but few can ticulate the principle of syntax that this utterance violates.

Education often serves to teach the child to make implicit theories explicit. ut some theories that the learner is asked to master explicitly conflict with ‹isting implicit theories. For example, theories of mechanics developed ırly in life may interfere with the learning of formal theories of mechanics. ıfants, as we have seen, initiate hand movements to intercept objects on e implicit assumption that they will continue along curvilinear trajectories. IcCloskey (1983) suggests that such extrapolations could contribute to a edieval impetus theory of moving objects. There is evidence that students ven in high school and college have trouble assimilating Newtonian meıanics because they convert what they are taught into something like the ıpetus theory. For example, a student who had completed college physics, ked to define momentum, replied: ". . . A combination of the velocity ıd the mass of an object. It's something that keeps a body moving." learly, he had not grasped the concept of inertia, but like the child or the edieval physicist, considered that some force is always required to keep ı object moving at constant velocity.

The spontaneous development of early knowledge structures makes it ›ssible for infants and very young children to acquire rapidly a functional ıderstanding of the world. Yet, these early theories can be two-edged vords; they may sometimes impede children's understanding of explicit eories encountered in the context of formal schooling. Knowing this, we are in a better position to understand some of the problems children have in school.

The Expansion of Strategic Powers

Young children have considerable strategic skill. Still, they have a long way to go in meeting the demands of literacy. Three-year-olds keeping alive their memory of a hidden toy seem to grasp the rudiments of rehearsal, but this does not mean that they know how to rehearse in a manner that would assist them in learning to spell, or remember historical facts or complex logical or mathematical relations. Gradual refinement and tuning of skills, together with a growing understanding of their function and range of utility, typifies the evolution of many school-relevant learning strategies. An example is skill at learning word lists. Two-year-olds display primitive precursors of rehearsal in their attempts to maintain memory of an object by naming, pointing, or eye fixation (DeLoache et al., 1985). By five years of age, children attempt to name (label) some of the items in a set some of the time (Flavell et al., 1966). Labeling and rote repetition of single items become well established during the early grade-school years (Craik and Watkins, 1973). With increasing sophistication, children then begin to place more items in their rehearsal sets, engaging in "cumulative rehearsal." During the later primary and early secondary years there is continual refinement of cumulative rehearsal, such as coordinating acquisition and retrieval components and increasingly attending to the size and composition of rehearsal sets (Belmont and Butterfield, 1977). Adolescents use elaborated rehearsal: they become increasingly sensitive to the presence of conceptual organization in the to-be-remembered list and capitalize on this inherent structure whenever possible (Ornstein and Naus, 1978), a development necessary to moving from rehearsal of lists of items and paired associates (as in spelling and foreign language learning) to the learning of whole segments of text. Adequate rehearsal strategies for studying do not appear until well into the high school years and are not perfected even by college students (Brown et al., 1983).

Memory strategies are not the only forms of school-related learning that evolve gradually. Literacy also demands skills of exposition and communication far beyond those expected of the preliterate child. Although young children can take their listeners' knowledge, perspective, and communicative competence into account when attempting to relay a simple message, schools demand much greater sophistication. The student is often required to communicate hazily understood material to an audience that does not require that the student communicate in writing to an unseen, often unknown audience, remote in time and space.

It is useful then to think of competence in terms of bandwidths, the lower end defined by the spontaneous learning of early childhood, the upper end defined by the ever increasing demands of the literate and technological society served by the schools (Brown and Reeve, 1985). Early competence emerges in hospitable contexts that match well with the child's knowledge, interests, and goals: here we see the "tireless explorer" and the "knowledge seeker" (Chukovsky, 1971) in action, the "little scientist" coming to understand his world. In schools, however, the goals and contexts of learning cannot always be of the child's choosing. The goal of learning through spontaneous discovery cannot always be maintained, and students must acquire skills of learning for learning's sake. By its very nature, much of schooling must be divorced from the simple, readily understandable goals of play or work (Bruner, 1972). Formal learning demands that students acquire knowledge without context, and even the preferred structuring of knowledge in temporal, spatial scripts, or story form, must be waived in favor of academic forms of organization by hierarchy and taxonomy (Mandler, 1983). It should not be surprising that many children's natural learning proclivities are overwhelmed by the task of acquiring large amounts of decontextualized material, organized in nonpreferred modes, with demands for precision and processing capacity greater than is the case in everyday life (Bartlett, 1958).

School learners not only must acquire knowledge in specific domains, such as science and history, but they must also "learn how to learn," developing routines for studying in general. More than ever before, schools must equip people to deal with facts that they will encounter only after they leave school. In a scientific and technological society based on an increasingly complex and rapidly changing information base, a productive member of society must be able to acquire new facts, critically evaluate them, and adapt to their implications. Schools need to develop intelligent novices (Brown et al., 1983), those who, although they may not possess the background knowledge needed in a new field, know how to go about gaining that knowledge.

Formal and Informal Teaching

It is not only the type of material to be learned that shifts in the school setting. There is also a substantial change in the teaching procedures compared with informal settings such as homes, preschools, or special interest clubs. In many cultures children are initiated into adult work activities and literacy events without explicit formal instruction. Opportunities for learning occur primarily in group settings where the adults are primarily responsibl for getting the task done. Children participate initially as spectators, later a novices responsible for a little of the work. As children become more experienced and capable of performing more complex tasks that have been demonstrated by adults time and time again, they are gradually ceded greate responsibility. Adults and children come to share the work, with a child takin initiative and an adult correcting and guiding where the child falters. Eventually, the adult gives the child the major active role and adopts the stance o a supportive audience. In these systems of tutelage, learning proceeds at th child's own pace; participation is expected only at a level the child can handle or a little beyond, thereby presenting a comfortable challenge.

The main features of natural learning situations are thus quite differen from formal schooling. In informal learning situations the group has responsibility for getting the job done, or at least an illusion of collectiv responsibility is maintained. The child joins in, often by self-initiative o with seemingly little adult pressure. Everyone has the same, clearly define agenda. The adult (parent, expert, master craftsman) models the matur behavior and guides the novice to increasingly more mature participation There is rarely individual testing; indeed, it is difficult to measure the child' individual contribution because everyone is participating at the same time The child performs within a limited zone of competence and is rarely calle upon to perform beyond capacity; the group does not expose the child' ignorance, but jointly benefits from the child's increasing competence.

Formal schooling makes a sharper distinction between expert and novic status, placing a heavy premium on individual performance. The teache the knowledge giver, demonstrates, lectures, and directs the children, th knowledge receivers. Some of these verbal interchanges are unfamiliar. Fo example, a typical classroom dialogue is the question-answer-evaluatio sequence (Mehan, 1979); the teacher addresses a question to a child, receives a response, and evaluates it explicitly (''good'') or implicitly (b ignoring the incorrect answer and calling on another child). Typicall children are called upon to perform independently, often when not read to do so. They do not always know in advance what they will be responsibl for and the questions are not always at the appropriate level. They run th risk of failing publicly. Understanding of appropriate turn-taking rituals i acquired slowly by some children, and such practices are even contrary t the approved social patterns of some cultures (Au, 1980; Heath, 1981).

Learned Academic Helplessness

Faced with challenges to their evolving partial theories, preferred learnin styles, and modes of interaction, a sizable minority of children react t schooling by becoming somewhat passive learners. Habitual failure in academic settings erodes their feelings of personal competence. The additional burden of repeated evaluation and labeling that accompanies continued failure is even more damaging. Such children develop quite devastating diagnoses of their own capabilities, readily describing themselves as "dumb," "not good at school things," "too stupid to read," etc. They come to question their personal efficacy (Bandura, 1980) in school settings. Children who view themselves as inadequate in school, as nonstarters in the academic race, often develop compensatory coping strategies to preserve their feelings of self-worth in what they view as the less-than-hospitable environment of the classroom.

Negative conceptions of one's prognosis for school success lead at best to defensive "passing," "coping," or "managing" (Goffman, 1963). Coping strategies include systematic devaluation of academic tasks and goals and the justification of lack of effort, i.e., "who needs to read anyways." Passing and managing tactics can be perfected so that the wily child avoids occasions of challenge. Threatening tests can be avoided if other children will cover; teachers will avoid embarrassment by not calling on the weaker child (Cole and Traupmann, 1980). All these ploys serve to defend against damaging expositions, attributions of failure, and further erosion of self-efficacy. These defenses are also formidable barriers to learning. Orienting one's attention and effort in school to minimizing demonstrations of failure rather than actively seeking occasions for acquiring new knowledge may be a realistic reaction to repeated obstacles, but it is not conducive to new learning.

Failure-oriented children typically display a pattern of learned helplessness in the face of obstacles or errors (Seligman et al., 1971). This pattern also increases negative feelings and further deflates the prognosis for success. There is a concomitant degradation of learning strategies. Failure-oriented children attribute their errors to lack of ability and often view temporary failure as an indication of a stable, generalized incompetence ("I'm dumb."). Helpless children question their ability in the face of obstacles, perceiving past successes to be few and irrelevant and future effort to be futile (Dweck and Bempechat, 1983).

In contrast, mastery-oriented children treat obstacles as challenges to be overcome by perfecting one's learning strategies; they do not attribute a temporary setback to personal shortcomings. Their verbalizations following failure often consist of positive self-instruction: "Slow down," "try new tactics," "evaluate the task more systematically." Dweck and Bempechat (1983) argue that these different reactions to academic difficulties reflect whether the child conceives tasks in terms of performance goals, where competence is to be evaluated and perhaps found wanting, or learning goals, where an opportunity exists to acquire new competences. Performance-goal children feel that they have been successful when they ''don't make mistakes,'' ''get easy work,'' etc. whereas learning-goal children feel successful when they master a new skill.

Reawakening the Active Learner

To be successful, interventions with passive learners must reinstill the confidence necessary for self-directed learning. Wertime (1979) has argued that many students need help to increase their ''courage spans,'' enabling them to treat failures as false starts or blind alleys that can be overcome and to regard errors as useful information. Students need to tolerate ambiguity, evaluate and judge information, and seek disconfirming evidence—in short, become critics and especially self-critics (Binet, 1909; Brown, 1985). But this criticism must be constructive, mastery-oriented self-guidance rather than self-derogation.

To end on a optimistic note we will illustrate two methods that have achieved some success at acclimatizing children to formal learning settings: (a) avoiding initial failures by adapting early school experiences to the prior competence of the entering child; and (b) lessening the gap between informal and formal teaching styles.

An excellent example of matching classrooms to homes is Heath's work with poor black Appalachian kindergarten children entering classrooms of white middle-class teachers (Heath, 1981). Heath found systematic differences between questioning behavior in the black and white communities she studied, particularly a mismatch between classroom questioning routines and spontaneous questioning activities in black preschoolers' environments. A common classroom routine is the ''known-answer'' question. Teachers routinely call on children to answer questions in order to display the children's knowledge rather than to provide information that the teacher does not have, which is the more familiar purpose of a question. These classroom questioning patterns do not map well into the earlier experiences of many children who lack informal exposure to academic language games.

At the beginning of the study Heath found that teachers were bewildered by the lack of responsiveness of their black pupils. For example: ''They don't seem to be able to answer even the simplest questions.'' ''I would almost think they have a hearing problem; it's as if they don't hear me ask a question.'' ''I sometimes feel that when I look at them and ask a question, I'm staring at a wall I can't break through'' (Heath, 1981:108).

Heath shared with teachers her documentation of the types of preschool questioning these children were familiar with, such as metaphoric and narrative sequences, and encouraged them to engineer settings that evoked the children's competence in the familiar format. Having practiced familiar questioning rituals, the teachers were then able to introduce the unfamiliar known-answer routines with great success. Another case of easing the transition to formal schooling, by capitalizing on the children's strengths rather than exposing their weaknesses, is the remarkable gains in reading achievement shown by Native Hawaiian (Polynesian) children after reading lessons were set in the context of a familiar Hawaiian interactive game, "talk-story" (Au, 1980).

Another successful intervention ploy is to lessen the gap between informal and formal learning settings. As we have seen, natural tutoring involves modeling on the part of the teacher and a gradual transfer of responsibility to the novices when and if they are ready to take control of their own learning. Instructional routines that mimic natural tutoring sessions are proving quite successful. For example, junior high school "passive" learners with depressed reading comprehension scores were moved from traditional instruction to a reciprocal teaching environment based on theories of natural tutoring. In reciprocal teaching, students of varying levels of competence and an adult teacher take turns "being the teacher," that is, leading a dialogue on a segment of text they are jointly attempting to understand and remember. The teacher responsible for a particular segment of text leads the ensuing dialogue by stating the gist in his or her own words, posing a question, clarifying any misunderstandings, and predicting what might happen next. All of these activities are part of a natural dialogue between the adult teacher and students. If a student has difficulty with any component of the dialogue, the teacher provides modeling and feedback at the student's current level, gradually leading each student to independent competence. Examples of such gradual transfer of responsibility can be found in Palincsar and Brown (1984).

Reciprocal teaching is based on certain central principles of effective learning: (1) the teacher models the desired comprehension activities, thereby making underlying processes overt, explicit, and concrete; (2) the teacher demonstrates the activities in appropriate contexts, not as isolated decontextualized skills; (3) the students are fully informed of the need for strategic intervention and the range of utility of a particular strategy; (4) the students see immediately that the use of strategies works for them; (5) the responsibility for the comprehension activities is transferred to the students as soon as they can take charge of their own learning; (6) this transfer of responsibility is gradual, presenting students with a comfortable challenge; and (7) feedback is tailored to the students' existing levels, encouraging them to progress one more step toward competence.

The reciprocal teaching procedure involves continuous trial and error on the part of the student, coupled with continuous adjustment on the part of the teacher to the student's current competence. Through inte with the supportive teacher and their more knowledgeable pe students are led to perform at an increasingly more mature level times this progress is fast, sometimes slow, but, irrespective of t the teacher provides an opportunity for the students to respond at a challenging level. As the students master one level of involvem teacher increases his demands so that the students are graduall upon to adopt the adult role fully and independently. The teac fades into the background as the students take charge of their own l from texts.

The results of the reciprocal teaching intervention with junior high ers were dramatic. The students improved their ability to clarify, summarize, and ask questions. Consider the quality of the summari seventh-grade students initially produced summaries ranked inadequ by the standards set by fifth graders. At the end of two weeks reciprocal teaching sessions, they were able to produce quite ac inventions, i.e., summaries, in their own words, of the gist of a p dialogue. A predominance of inventions characterizes the untrain marization performance of college freshmen (Brown and Day, 1983 guided instruction had taken these failing seventh graders to a competence far beyond that typical for their peers. Furthermore, t became able to assume the role of teacher, producing their own q and summaries and evaluating those of others. In addition, the significant improvements in independent performance on laborator room, and standardized tests of comprehension. But perhaps mor tantly, the children's feelings of personal competence and control i dramatically. Allowed to take charge of the dialogues, and even t advanced students, these "failing" students increased their courag as their purely cognitive skills. Success bred positive expectatio teachers and improved students' personal "efficacy," i.e., the co to employ active learning strategies in the belief that they will w

It is important to note that mimicking natural tutoring styles ha a successful instructional technique in areas other than reading: comprehension (Brown and Palincsar, in press), writing (Apple Langer, 1983; Scardamalia, 1984), storytelling (McNamee, 1981), (Frase and Schwartz, 1976), and problem solving (Bloom and Brode have all responded well to reciprocal instruction strategies. In ad is not only teachers who can serve as the agent of change but also (Ninio and Bruner, 1978; Saxe et al., 1984; Scollon, 1976; Wertsch peers (Bloom and Broder, 1950; Whimbey and Lochhead, 1982), somewhat intelligent computer tutors (Brown et al., 1982; Heller a gate, 1984; Lesgold and Reif, 1983). The concept of expert sca the gradually guided transfer of learning responsibility from an expert to a novice, has wide applicability as an instructional philosophy.

Summary

Recognition of children's natural competence, both in terms of strategic rules and knowledge, is having a profound effect on instructional theory. Structured instruction, however, is necessary for the child to go beyond imprecise, and sometimes erroneous, implicit theories and to acquire the precise, explicit theories that constitute formal knowledge. Through the intervention of certain forms of formal schooling, children are turned into *routine* school *experts* (Hatano, 1982), able to perform, more and more efficiently, the procedures taught and practiced in schools.

One problem, however, is that routine expertise can lead to the acquisition of "inert knowledge" (Whitehead, 1916), acquired by rote learning and practice but rarely used flexibly and creatively. Educational systems that promote *adaptive expertise* (Hatano, 1982), whereby students come to understand, challenge, and flexibly apply their knowledge, depend on maintaining the active thirst for knowledge that the preschool child brings initially into settings of formal education. The more we learn about the knowledge structures that children bring to school and the instructional practices that foster their natural proclivities to build and refine theories, the more able we will be to design instructional modes that promote adaptive expertise rather than the acquisition of inert knowledge.

CONCLUSION

In this chapter we have concentrated on an apparent paradox concerning the cognitive competence of children. Recent research with infants and very young children suggests that they know far more about their world initially, and develop this understanding more rapidly, than was previously supposed. However, topical consternation over the putatively increasing incompetence of school-aged children in academic settings stands in sharp contrast to these claims of early ingenuity.

In the first part of the chapter, we discussed the necessity of granting complex cognitive structures to the young human mind. This breaking away from an empiricist account of human thought took its impetus from sweeping changes in psychological theory pioneered notably by Chomsky, the Gibsons, and Piaget. Buttressing these theoretical claims is a body of contemporary research gleaned from a variety of ingenious techniques that make it increasingly feasible to interrogate infants. The outcome of a painstaking set of inquiries is a window through which we can view the young cognitive world, a window that is only beginning to open.

We now know that infants are sensitive to certain principles of mc early in life; that they garner multisensory and multimodal informati the nature of objects; that they endow objects with properties of rigi solidity; and that they possess rudimentary theories of categories nizing properties of sets of nonidentical objects, including nume property of sets divorced from any description of the objects ther Implicit principles of causality, numerosity, etc. guide the develop such knowledge at a rapid pace during the preschool years, a tim which children are busily engaged in exploring their environmen acterized as "tireless explorers," they invent primitive but ser comprehension, learning, and memory strategies, and create and t tinously evolving theories to breathe meaning into their physical ar world.

The pace of this development seems to slow down during th years, but this may be because children's competence is increasingly in the light of their performance on academic tests. Learning in differs from natural learning in that others are in charge of what learned, others control the timetable, and students must develop and skill in learning for learning's sake so that they can intentio about acquiring large bodies of decontextualized knowledge.

In an increasingly complex and rapidly changing technological more than ever before, students must be equipped to acquire ne mation, critically evaluate it, and adapt to its implications. They m to waive their imprecise theories in favor of the precise, explic encompassing theories that constitute formal knowledge. Profoun change of this magnitude comes at a cost that many may be rel pay without a supportive academic environment. In the latter pa chapter, we discussed innovative pedagogical procedures that serve tain and bolster the child's natural curiosity and theory-building ca In the exploitation of such techniques lies hope for solving the pa early competence and later academic crisis.

REFERENCES

Applebee, A.N., and Langer, J.A.
1983 Instructional scaffolding: reading and writing as natural language activities. *Arts* 60:168–175.

Aslin, R.N., Pisoni, D.B., and Jusczyk, P.W.
1983 Auditory development and speech perception in infancy. In M.M. Hait Campos, eds., *Handbook of Child Psychology*. Vol. 2. *Infancy and Dev Psychobiology*. New York: John Wiley and Sons.

Au, K.H.

1980 A Test of the Social Organizational Hypothesis: Relationships Between Participation Structures and Learning to Read. Unpublished doctoral dissertation, University of Illinois.

Baillargeon, R., Spelke, E.S., and Wasserman, S.

In press Object permanence in 5-month-old infants. *Cognition.*

Bandura, A.

1980 Self-referent thought: the development of self-efficacy. In J.H. Flavell and L.D. Ross, eds., *Development of Social Cognition.* Hillsdale, N.J.: Lawrence Erlbaum Associates.

Bartlett, F.C.

1958 *Thinking: An Experimental and Social Study.* New York: Basic Books.

Belmont, J.M., and Butterfield, E.C.

1977 The instructional approach to developmental cognitive research. In R.V. Kail, Jr., and J.W. Hagen, eds., *Perspectives on the Development of Memory and Cognition.* Hillsdale, N.J.: Lawrence Erlbaum Associates.

Binet, A.

1909 *Les Idées Modernes sur les Enfants.* Paris: Ernst Flamarion.

Bjork, E.L., and Cummings, E.M.

1984 Infant search errors: stage of concept development or stage of memory development. *Memory & Cognition* 12(1):1–192.

Bloom, B.S., and Broder, L.J.

1950 *Problem-Solving Processes of College Students.* Chicago: University of Chicago Press.

Braine, M.D.S., and Rumain, B.

1983 Logical reasoning. In J.H. Flavell and E.M. Markman, eds., *Handbook of Child Development* (4th ed.). Vol. 3. New York: John Wiley and Sons.

Brown, A.L.

1975 The development of memory: knowing, knowing about knowing, and knowing how to know. In H.W. Reese, ed., *Advances in Child Development and Behavior.* Vol. 10. New York: Academic Press.

1985 Mental orthopedics: a conversation with Alfred Binet. In S. Chipman, J. Segal, and R. Glaser, eds., *Thinking and Learning Skills: Current Research and Open Questions.* Vol. 2. Hillsdale, N.J.: Lawrence Erlbaum Associates.

Brown, A.L., and Day, J.D.

1983 Macrorules for summarizing texts: the development of expertise. *Journal of Verbal Learning and Verbal Behavior* 22:1–14.

Brown, A.L., and DeLoache, J.S.

1978 Skills, plans, and self-regulation. In R. Siegler, ed., *Children's Thinking: What Develops?* Hillsdale, N.J.: Lawrence Erlbaum Associates.

Brown, A.L., and Palincsar, A.S.

In press Reciprocal teaching of comprehension strategies: a natural history of one program for enhancing learning. In J. Borkowski and J.D. Day, eds., *Intelligence and Cognition in Special Children: Comparative Studies of Giftedness, Mental Retardation, and Learning Disabilities.* New York: Ablex.

Brown, A.L., and Reeve, R.A.

1985 Bandwidths of competence: the role of supportive contexts in learning and development. In L.S. Liben and D.H. Feldman, eds., *Development and Learning: Conflicts or Congruence?* Hillsdale, N.J.: Lawrence Erlbaum Associates.

Brown, A.L., Bransford, J.D., Ferrara, R.A., and Campione, J.C.

1983 Learning, remembering, and understanding. In J.H. Flavell and E.M. Markman, eds., *Handbook of Child Psychology* (4th ed.). Vol. 3. *Cognitive Development.* New York: John Wiley and Sons.

Brown, J.S., and VanLehn, K.

1982 Toward a generative theory of "bugs." In T.P. Carpenter, J.M. Moser, and T.A. Romberg, eds., *Addition and Subtraction: A Cognitive Perspective.* Hillsdale, N.J.: Lawrence Erlbaum Associates.

Brown, J.S., Burton, R.R., and deKleer, J.

1982 Pedagogical, natural language and knowledge engineering techniques in SOPHIE I, II and III. In D. Sleeman and J.S. Brown, eds., *Intelligent Tutoring Systems.* New York: Academic Press.

Bruner, J.S.

1972 Nature and uses of immaturity. *American Psychologist* 27:687–708.

Bullock, M.

In press Preschoolers' understanding of causal connections. *British Journal of Developmental Psychology.*

Bullock, M., Gelman, R., and Baillargeon, R.

1982 The development of causal reasoning. In J. Friedman, ed., *The Developmental Psychology of Time.* New York: Academic Press.

Carey, S.

1985a Are children fundamentally different kinds of thinkers and learners than adults? In S. Chipman, J. Segal, and R. Glaser, eds., *Thinking and Learning Skills.* Hillsdale, N.J.: Lawrence Erlbaum Associates.

1985b *Conceptual Change in Childhood.* Cambridge, Mass.: MIT Press/Bradford.

Chomsky, N.

1957 *Syntactic Structures.* The Hague: Mouton.

Chukovsky, K.

1971 *From Two to Five* (revised ed.). Berkeley: University of California Press. Translation by Miriam Morton, 1963.

Cole, M., and Traupmann, K.

1980 Comparative cognitive research: learning from a learning disabled child. In A. Collins, ed., *Minnesota Symposium on Child Development.* Hillsdale, N.J.: Lawrence Erlbaum Associates.

Craik, F.I.M., and Watkins, M.J.

1973 The role of rehearsal in short-term memory. *Journal of Verbal Learning and Verbal Behavior* 12:599–607.

DeLoache, J.S., Cassidy, D.J., and Brown, A.L.

1985 Precursors of mnemonic strategies in very young children's memory. *Child Development* 56:125–137.

DeLoache, J.S., Sugarman, S., and Brown, A.L.

In press The development of error correction strategies in young children's manipulative play. *Child Development.*

Donaldson, M.A.

1978 *Children's Minds.* New York: Norton.

Dweck, C.S., and Bempechat, J.

1983 Children's theories of intelligence: consequences for learning. In S.G. Paris, G.M. Olson, and H.W. Stevenson, eds., *Learning and Motivation in the Classroom.* Hillsdale, N.J.: Lawrence Erlbaum Associates.

Eimas, P.D., Siqueland, E.R., Jusczky, P.W., and Vigorito, J.

1971 Speech perception in infants. *Science* 171:303–306.

Evans, D.W.

1983 Understanding Zero and Infinity in the Early School Years. Unpublished doctoral dissertation. University of Pennsylvania, Philadelphia.

Fantz, R.L.

1961 The origins of form perception. *Scientific American* 204:66–72.

1966 Pattern discrimination and selective attention as determinants of perceptual development from birth. In A.H. Kidd and J.L. Rivoire, eds., *Perceptual Development in Children*. New York: International Universities Press.

Field, D.

1978 How children in ESN schools in London learn conservation skills. In G.L. Lubin, M.K. Poulsene, J.F. Magary, and M. Soto-McAlister, eds., *Piagetian Theory and Its Implications for the Helping Professions*. Proceedings of the Seventh Inter-disciplinary Conference. Vol. 2. Los Angeles: University of Southern California Press.

Field, J., Muir, D., Pilon, R., Sinclair, M., and Dodwell, P.C.

1980 Infants' orientation to lateral sounds from birth to three months. *Child Development* 50:295–298.

Flavell, J.H.

1970a Concept development. In P.H. Mussen, ed., *Carmichael's Manual of Child Psychology*. Vol. 1. New York: John Wiley and Sons.

1970b Developmental studies of mediated memory. In H.W. Reese and L.P. Lipsitt, eds., *Advances in Child Development and Behavior*. Vol. 5. New York: Academic Press.

Flavell, J.H., Beach, D.H., and Chinsky, J.M.

1966 Spontaneous verbal rehearsal in memory tasks as a function of age. *Child Development* 37:283–299.

Frase, L.T., and Schwartz, B.J.

1975 The effect of question production and answering on prose recall. *Journal of Educational Psychology* 62:628–635.

Gelman, R.

1977 How young children reason about small numbers. In N.J. Castellan, D.B. Pisoni, and G.R. Potts, eds., *Cognitive Theory*. Vol. 2. Hillsdale, N.J.: Lawrence Erlbaum Associates.

1978 Cognitive development. *Annual Review of Psychology* 29:297–332.

Gelman, R., and Gallistel, C.R.

1978 *The Child's Understanding of Number*. Cambridge, Mass.: Harvard University Press.

Gelman, R., and Meck, E.

1983 Preschoolers' counting: principles before skill. *Cognition* 13:343–359.

Gelman, R., Spelke, E.S., and Meck, E.

1983 What preschoolers know about animate and inanimate objects. In D. Rogers and J.A. Sloboda, eds., *The Acquisition of Symbolic Skills*. London: Plenum.

Gibson, E.J., and Spelke, E.S.

1983 The development of perception. In J.H. Flavell and E.M. Markman, eds., *Handbook of Child Psychology* (4th ed.). Vol. 3. *Cognitive Development*. New York: John Wiley and Sons.

Gibson, E.J., and Walker, A.S.

1984 Intermodal perception of substance. *Child Development* 55:453–460.

Goffman, E.

1963 *Stigma: Notes on the Management of Spoiled Identity*. Englewood Cliffs, N.J.: Prentice-Hall.

Greeno, J.G., Riley, M.S., and Gelman, R.

1984 Conceptual competence and children's counting. *Cognitive Psychology* 16:94–143.

Groen, G.J., and Resnick, L.B.

1977 Can preschool children invent addition algorithms? *Journal of Educational Psychology* 69:645–652.

Harris, P.L.

1983 Infant cognition. In M. Miltaith and S.S. Campos, eds., *Handbook of Child F* (4th ed.). Vol. 2. *Infancy and Developmental Psychobiology*. New York: J and Sons.

Hatano, G.

1982 Cognitive consequences of practice in culture specific procedural skills. *The Newsletter of the Laboratory of Comparative Human Cognition* 4:15–18.

Heath, S.B.

1981 Questioning at home and at school: a comparative study. In G. Spindler, e *Ethnography: Educational Anthropology in Action*. New York: Holt, Ri Winston.

Heller, J.I., and Hungate, H.N.

1984 Computer-based Expert Scaffolding. Unpublished manuscript. University nia, Berkeley.

Hood, L., and Bloom, L.

1979 What, when and how about why: a longitudinal study of early expressions of *Monographs of the Society for Research in Child Development* 44 (Serial No. 6).

Kagan, J., Kearsley, R.B., and Zelaso, P.R.

1978 *Infancy: Its Place in Human Development*. Cambridge, Mass.: Harvard Press.

Kail, R.V., Jr., and Hagen, J.W., eds.

1977 *Perspectives on the Development of Memory and Cognition*. Hillsdale, N.J.: Erlbaum Associates.

Kalnins, I.V., and Bruner, J.S.

1973 The coordination of visual observation and instrumental behavior in earl *Perception* 2:307–314.

Karmiloff-Smith, A.

1979 Problem solving construction and representations of closed railway circuits *of Psychology* 47:37–59.

1985 Children's problem solving. In M. Lamb, A.L. Brown, and B. Rogoff, eds. *in Developmental Psychology*. Vol. 3. Hillsdale, N.J.: Lawrence Erlbaum A

Karmiloff-Smith, A., and Inhelder, B.

1974–1975 If you want to get ahead, get a theory. *Cognition*, 3, pp. 195–212.

Keil, F.C.

1979 *Semantic and Conceptual Development*. Cambridge, Mass.: Harvard Unive

1981 Constraints on knowledge and cognitive development. *Psychological Revie* 227.

Kessen, W., Salapateck, P.H., and Haith, M.M.

1972 The visual response of the human newborn to linear contour. *Journal of Ex Child Psychology* 13:19–20.

Klahr, D., and Wallace, J.G.

1976 *Cognitive Development, an Information Processing View*. Hillsdale, N.J. Erlbaum Associates.

Koslowski, B., Spilton, D., and Snipper, A.

1981 Children's beliefs about instances of mechanical and electrical causation. *Applied Developmental Psychology* 2:189–210.

Laboratory of Comparative Human Cognition

1983 Culture and cognitive development. In P.H. Mussen, ed., *Handbook of Child* (4th ed.). Vol. 1. *History, Theory, and Methods*. New York: John Wiley ar

andau, B., Gleitman, H., and Spelke, E.S.
1981 Spatial knowledge and geometric representation in a child blind from birth. *Science* 213:1275–1278.

empers, J.D., Flavell, E.R., and Flavell, J.H.
1977 The development in very young children of tacit knowledge concerning visual perception. *Genetic Psychology Monographs* 95:3–53.

esgold, A.M., and Reif, F.
1983 *Computers in Education: Realizing the Potential.* (Report of 1982 Pittsburgh Research Conference, U.S. Department of Education). Washington, D.C.: U.S. Government Printing Office.

indvall, C.M., and Ibarra, C.G.
1981 Incorrect Procedures Used by Primary Grade School Pupils in Solving Open Addition and Subtraction Sentences. Manuscript, Learning, Research and Development Center, University of Pittsburgh, Pittsburgh, Pa.

andler, J.M.
1983 Representation. In J.H. Flavell and E.M. Markman, eds., *Handbook of Child Psychology*, (4th ed.). Vol. 3. *Cognitive Development*, pp. 420–494. New York: John Wiley and Sons.

arkman, E.M.
1981 Two different principles of conceptual organization. In M.E. Lambe and A.L. Brown, eds., *Advances in Developmental Psychology.* Vol. 1. Hillsdale, N.J.: Lawrence Erlbaum Associates.

cCloskey, M.
1983 Naive theories of motion. In D. Gentner and A.L. Stevens, eds., *Mental Models.* Hillsdale, N.J.: Lawrence Erlbaum Associates.

cNamee, G.D.
1981 Social Origins of Narrative Skills. Paper presented at April meeting of the Society for Research in Child Development, Boston.

ehan, H.
1979 *Learning Lessons: Social Organization in the Classroom.* Cambridge, Mass.: Harvard University Press.

eltzoff, A.N., and Borton, R.
1979 Intermodal matching by human neonates. *Nature* 282:403–404.

endelson, M.J., and Haith, M.M.
1976 The relation between audition and vision in the human newborn infant. *Monographs of the Society for Research in Child Development* 41 (4, Serial No. 167).

elson, K., and Gruendel, J.
1981 Generalized event representations: basic building blocks of cognitive development. In M.E. Lambe and A.L. Brown, eds., *Advances in Developmental Psychology.* Vol. 1, pp. 131–158. Hillsdale, N.J.: Lawrence Erlbaum Associates.

inio, A., and Bruner, J.S.
1978 The achievement and antecedents of labelling. *Journal of Child Language* 5:1–15.

rnstein, P.A., and Naus, M.J.
1978 Rehearsal processes in children's memory. In P.A. Ornstein, ed., *Memory Development in Children.* Hillsdale, N.J.: Lawrence Erlbaum Associates.

alincsar, A.S., and Brown, A.L.
1984 Reciprocal teaching of comprehension-fostering and monitoring activities. *Cognition and Instruction* 1:(2)117–175.

iaget, J.
1954 *The Construction of Reality in the Child.* New York: Basic Books.
1980 *Experiments in Contradiction.* Chicago and London: University of Chicago Press.

Premack, D.
1976 *Intelligence in Ape and Man*. Hillsdale, N.J.: Lawrence Erlbaum Asso

Rogoff, B., and Wertsch, J.V., eds.
1984 *Children's Learning in the "Zone of Proximal Development."* San Fran Bass.

Rosch, E., Mervis, C.B., Gray, W.D., Boyes-Braem, P., and Johnson, D.M.
1976 Basic objects in natural categories. *Cognitive Psychology* 8:382–439.

Ross, G.S.
1980 Categorization in 1- to 2-year-olds. *Developmental Psychology* 16:391-

Saxe, G.B., Gearhart, M., and Guberman, S.R.
1984 The social organization of early number development. In B. Rogoff an eds., *Children's Learning in the "Zone of Proximal Development."* S Jossey-Bass.

Scardamalia, M.
1984 Knowledge Telling and Knowledge Transforming in Written Compositi sented at the meeting of the American Educational Research Association,

Scollon, R.
1976 *Conversations with a One-Year-Old*. Honolulu: University Press of Ha

Seligman, M.E.P., Maier, S.F., and Solomon, F.L.
1971 Unpredictable and uncontrollable aversive events. In S.R. Brush, ed., *A ditioning and Learning*. New York: Academic Press.

Shatz, M.
1978 The relationship between cognitive processes and the development of co skills. In B. Keasey, ed., *Nebraska Symposium on Motivation*. Lincol of Nebraska Press.

Shultz, T.R.
1982 Rules of causal attribution. *Monographs of the Society for Research in opment* 47 (1, Serial No. 194).

Siegler, R.S., and Robinson, M.
1982 The development of numerical understandings. In H.W. Reese, ed., *Adva Development and Behavior*. Vol. 16. New York: Academic Press.

Simon, H.A.
1972 On the development of the processor. Pp. 3–22 in S. Farnham-Diggory *mation Processing in Children*. New York: Academic Press.

Sophian, C.
1984 Developing search skills in infancy and early childhood. In C. Sophian, *of Cognitive Skills*. Hillsdale, N.J.: Lawrence Erlbaum Associates.

Spelke, E.S.
1976 Infants' intermodal perception of events. *Cognitive Psychology* 8:553–5

Starkey, P., and Cooper, R.G.
1980 Numerosity perception in human infants. *Science* 210:1033–1035.

Starkey, P., Spelke, E.S., and Gelman, R.
1983 Detection of 1-1 correspondences by human infants. *Science* 222:79–81
In press Numerical abstraction by human infants. *Journal of Cognition*.

Stein, N.L.L., and Trabasso, T.
1982 What's in a story: an approach to comprehension and instruction. In R. *Advances in the Psychology of Instruction*. Vol. 2. Norwood, N.J.: Abl

Stevenson, H.W.
1970 Children's learning. In P.H. Mussen, ed., *Manual of Child Psychology*. York: John Wiley and Sons.

Strauss, M.S., and Curtis, L.E.

1981 Infant perception of numerosity. *Child Development* 52:1146–1152.

Tronick, E., Adamson, L., Wise, S., Als, H., and Brazelton, T.B.

1975 The Infant's Response to Entrapment Between Contradictory Messages in Face to Face Interaction. Paper presented at the meeting of the Society for Research in Child Development, Denver.

Von Hofsten, C.

1980 Predictive reaching for moving objects by human infants. *Journal of Experimental Child Psychology* 30:369–382.

Vygotsky, L.S.

1962 *Thought and Language*. Cambridge, Mass.: MIT Press.

Wanner, E., and Gleitman, L.R., eds.

1982 *Language Acquisition: The State of the Art*. Cambridge, England: Cambridge University Press.

Wellman, H.M., Ritter, R., and Flavell, J.H.

1975 Deliberate memory behavior in the delayed reactions of very young children. *Developmental Psychology* 11:780–787.

Wertheimer, M.

1961 Psychomotor coordination of auditory-visual space at birth. *Science* 134.

Wertime, R.

1979 Students, problems, and courage spans. In I. Lockhead and J. Clement, eds., *Cognitive Process Instruction: Research on Teaching Thinking Skill*. Philadelphia: Franklin Institute Press.

Wertsch, J.V.

1979 From social interaction to higher psychological processes: a clarification and application of Vygotsky's theory. *Human Development* 22:1–22.

Whimbey, A., and Lochhead, J.

1982 *Problem Solving and Comprehension*. Philadelphia: Franklin Institute Press.

White, S.H.

1965 Evidence for hierarchical arrangement of learning processes. In L.L. Lipsitt and C.C. Spiker, eds., *Advances in Child Development and Behavior*. Vol. 2. New York: Academic Press.

1970 The learning theory tradition for child psychology. In P.H. Mussen, ed., *Carmichael's Manual of Child Psychology*. Vol. 1. New York: John Wiley and Sons.

Whitehead, A.N.

1916 The Aims of Education. Presidential Address to the Mathematical Association of England.

Whorf, B.L.

1956 *Language, Thought and Reality*. John B. Carroll, ed. Cambridge, Mass.: MIT Press.

Some Developments in Research on Language Behavior

MICHAEL STUDDERT-KENNEDY

INTRODUCTION

Fifty years ago the study of language was largely a descriptive endeavor, grounded in the traditions of nineteenth century European philology. The object of study, as proposed by de Saussure in a famous course of lectures at the University of Geneva (1906–1911), was *langue*, language as a system, a cultural institution, rather than *parole*, language as spoken and heard by individuals. In 1933 historical linguists were describing and comparing the world's languages, tracing their family relations, and reconstructing the protolanguages from which they had sprung (Lehmann, 1973). Structural linguists were developing objective procedures for analyzing the sound patterns and syntax of a language, according to well-defined, systematic principles (e.g., Bloomfield, 1933). Students of dialect were applying such procedures to construct atlases of dialect geography (Kurath, 1939), while anthropological linguists were applying them to American Indian, African, Asian, Polynesian, and many other languages (Lehmann, 1973). The work goes on. From it we are coming to understand the origins of language diversity: not only how languages change over time and space but also how they and their dialects act as forces of social cohesion and differentiation (e.g., Labov, 1972).

However, the unfolding of the descriptive tradition and the development of new methods and theories in the field of sociolinguistics are not my concerns in this chapter. My concern, rather, is with a view of language that has emerged from a more diverse tradition. For like the taxonomic studies of Linnaeus in botany and of his followers in zoology, the great labor of language description and classification has provided the raw material for a broader science, stemming from the work of seventeenth century grammarians and of such nineteenth century figures as the German physicist Hermann von Helmholtz, the French neurologist Paul Broca, and the English phonetician Henry Sweet. The several strands that their works represent have come together over the past 30 to 40 years to form the basis of a new science of language, focusing on the individual, rather than on the social and cultural, linguistic system. Since the new focus is essentially biological, a biological analogy may be helpful. It is as though we shifted from describing and classifying the distinctive flight patterns of the world's eight or nine thousand species of birds to analyzing the basic principles of individual flight as they must be instantiated in the anatomy and physiology of every hummingbird and condor. Thus, this new science of language asks: What is language as a category of individual behavior? How does it differ from other systems of animal communication? What do individuals know when they know a language? What cognitive, perceptual, and motor capacities must they have to speak, hear, and understand a language? How do these capacities derive from their biophysical structures, that is, from human anatomy and physiology? What is the course of their ontogenetic development? And so on.

Such questions hardly fall within the province of a single discipline. The new field is markedly interdisciplinary and addresses questions of practical application as readily as questions of pure theory or knowledge. Linguistics, anthropology, psychology, biology, neuropsychology, neurology, and communications engineering all contribute to the field, and their research has implications for workers in many areas of social import: doctors and therapists treating stroke victims, surgeons operating on the brain, applied engineers working on human-machine communication, teachers of second languages, of reading, and of the deaf and otherwise language-handicapped.

The origins of the new science are an object lesson in the interplay between basic and applied research, and between research and theory. To understand this, we must begin by briefly examining the nature of language and the properties that make it unique as a system of communication.

The Structure of Language

If we compare language with other animal communication systems, we are struck by its breadth of reference. The signals of other animals form a closed set with specific, invariant meanings (Wilson, 1975). The ultrasonic squeaks of a young lemming denote alarm; the swinging steps and lifted tail of the male baboon summon his troop to follow; the "song" of the male white-crowned sparrow informs his fellows of his species, sex, local origin, personal identity, and readiness to breed or fight. Even the elaborate ''dance'' of the honeybee merely conveys information about the direction, distance, and quality of a nectar trove. But language can convey information about many more matters than these. In fact, it is the peculiar property of language to set no limit on the meanings it can carry.

How does language achieve this openness, or productivity? There are several key features to its design (Hockett, 1960). Here we note two. First, language is learned: it develops under the control of an open rather than a closed genetic program (Mayr, 1974). Transmission of the code from one generation to the next is therefore discontinuous; each individual recreates the system for himself. There is ample room here for creative variation—probably a central factor in the evolution of language and in the constant processes of change that all languages undergo (e.g., Kiparsky, 1968; Locke, 1983; Slobin, 1980). One incidental consequence of this freedom is that the universal properties of language (whatever they may be) are largely masked by the surface variety of the several thousand languages, and their many dialects, now spoken in the world.

Second, and more crucially, language has two hierarchically related levels of structure. One level, that of sound pattern, permits the growth of a large lexicon; the other level, that of syntax, permits the formation of an infinitely large set of utterances. A similar combinatorial principle underlies the structure of both levels.

Consider, first, the fact that a six-year-old, middle-class American child typically has a recognition vocabulary of some 8,000 root words, some 14,000 words in all (Templin, 1957). Most of these have been learned in the previous four years, at a rate of about five or six roots a day. As an adult, the child may come to have a vocabulary of well over 150,000 words (Seashore and Erickson, 1940). How is it possible to produce and perceive so many distinct signals?

The achievement evidently rests on the evolution in our hominid ancestors of a combinatorial principle by which a small set of meaningless elements (phonemes, or consonants and vowels) is repeatedly sampled, and the samples permuted, to form a very large set of meaningful elements (morphemes, words). Most languages have between 20 to 100 phonemes; English has about 40, depending on dialect. The phonemes themselves are formed from an even smaller set of movements, or gestures, made by jaw, lips, tongue, velum (soft palate), and larynx. Thus, the combinatorial principle was a biologically unique development that provided ''a kind of impedance match between an open-ended set of meaningful symbols and a decidedly limited set of signaling devices'' (Studdert-Kennedy and Lane, 1980; cf. Cooper, 1972; Liberman et al., 1967). We may note, incidentally, that a large lexicon is not peculiar to complex, literate societies: even so-called primitive human groups may deploy a considerable lexicon. For example, the Hanunoo, a stone age people of the Philippines, have nearly three thousand words for the flora and fauna of their world (Levi-Strauss, 1966).

Of course, a large lexicon is not a language. Many languages have relatively small lexicons, and in everyday speech we may draw habitually on no more than a few thousand words (Miller, 1951). To put words to linguistic use, we must combine them in particular ways. Every language has a set of rules and devices, its syntax, for grouping words into phrases, clauses, and sentences. Among the various devices that a language may use for predicating properties of objects and events, and for specifying their relations (who does what to whom) are word order and inflection (case, gender, and number affixes for nouns, pronouns, adjectives; person, tense, mood, and voice affixes for verbs). An important distinction is also made in all languages between open-class words with distinct meanings (nouns, verbs, adjectives, etc.) and closed-class or function words (conjunctions, articles, verbal auxiliaries, enclitics—e.g., the particle ''not'' in ''cannot'') that have no fixed meaning in themselves but serve the purely syntactic function of indicating relations between words in a sentence or sequence of sentences. Here again then, a combinatorial principle is invoked: a finite set of rules and devices is repeatedly sampled and applied to produce an infinite set of utterances.

I should note that many of the facts about language summarily described above are already framed from the new viewpoint that has developed in the past 40 years. Let us now turn back the clock and consider the early vicissitudes of three areas of applied research that contributed to this development.

Three Areas of Applied Research in Language

In the burst of technological enthusiasm that followed World War II, federal money flowed into three related areas of language study: automatic machine translation, automatic speech recognition, and automatic reading machines for the blind. A considerable research effort was mounted in all three areas during the late 1940s and early 1950s, but surprisingly little headway was made. The reason for this, as will become clear below, was that all three enterprises were launched under the shield of a behaviorist theory according to which complex behaviors could be properly described as chained sequences of stimuli and responses.

The initial assumption underlying attempts at machine translation was that this task entailed little more than transposing words (or morphemes) from one language into another, following a simple left-to-right sequence.

If this were so, we might store a sizable lexicon of matched Russian, say, and English words in a computer and execute translation by instructing the computer to type out the English counterpart of each Russian word typed in. Unfortunately, both semantic and syntactic stumbling blocks lie in the path. The range of meanings, literal and metaphorical, that one language assigns to a word (say, English *high*, as in "high mountain," "high pitch," "high hopes," "high horse," "high-stepping," and "high on drugs") may be quite different from the range assigned by another language; and the particular meaning to be assigned will be determined by context, that is, by meanings already assigned to some in principle unspecifiable sequence of preceding words. Moreover, the syntactic devices for grouping words into phrases, phrases into clauses, and clauses into sentences may be quite different in different languages. This is strikingly obvious when we compare a heavily inflected language, such as Russian, with a lightly inflected language with a more rigid word order, such as English. Oettinger (1972) amusingly illustrates the general difficulties with two simple sentences, immediately intelligible to an English speaker, but a source of knotty problems in both phrase structure and word meaning to a computer, programmed for left-to-right lexical assignment: *Time flies like an arrow*, and *Fruit flies like a banana*. From such observations, it gradually became clear that we would make little progress in machine translation without a deeper understanding of syntax and of its relation to meaning.

The initial assumption underlying attempts at automatic speech recognition was similar to that for machine translation and equally in error (cf. Reddy, 1975). The assumption was that the task entailed little more than specifying the invariant acoustic properties associated with each consonant and vowel, in a simple left-to-right sequence. One would then construct an acoustic filter to pass those properties but no others, and control the appropriate key on a printer by means of the output from each filter. Unfortunately, stumbling blocks lie in this path also. A large body of research has demonstrated that speech is not a simple left-to-right sequence of discrete and invariant alphabetic segments, such as we see on a printed page (e.g., Fant, 1962; Joos, 1948; Liberman et al., 1967). The reason for this, as we shall see shortly, is that we do not speak phoneme by phoneme, or even syllable by syllable. At each instant our articulators are engaged in executing patterns of movement that correspond to several neighboring phonemes, including those in neighboring syllables. The result of this shingled pattern of movement is, of course, a shingled pattern of sound. Even more extreme variation may be found when we examine the acoustic structure of the same syllable spoken with different stress or at different rates or by different speakers. From such observations it gradually became clear that we would make little progress in automatic speech recognition without a deeper understanding of how the acoustic structure of the speech signal specifies the linguistic structure of the message.

Finally, the initial assumption underlying attempts to construct a reading machine for the blind was closely related to that for automatic speech recognition and again in error (Cooper et al., 1984). A reading machine is a device that scans print and uses its contours to control an acoustic signal. It was supposed that, given an adequate device for optical recognition of letters on a page, one need only assign a distinctive auditory pattern to each letter, to be keyed by the optical reader and recorded on tape or played in real time to a listener—a sort of auditory Braille. Once again there were stumbling blocks, but this time they were perceptual. We normally speak and listen to English at a rate of some 150 words per minute (wpm), that is, roughly 5 to 6 syllables or 10 to 15 phonemes per second. Ten to 15 discrete sounds per second is close to the resolving power of the ear (20 elements per second merge perceptually into a low-pitched buzz). Not surprisingly, despite valiant and ingenious attempts to improve the acoustic array, even the most practiced listeners were unable to follow a substitute code at rates much beyond that of skilled Morse code receivers, namely some 10 to 15 words per minute—a rate intolerably slow for any extended use. From this work, it gradually became clear that the only acceptable output from a reading machine would be speech itself. This conclusion was one of many that spurred development of speech synthesis by artificial talking machines in following years (Cooper and Borst, 1952; Fant, 1973; Flanagan, 1983; Mattingly, 1968, 1974). The conclusion also raised theoretical questions. For example: Why can we successfully transpose speech into a visual alphabet, using another sensory modality, if we cannot successfully transpose it within its ''natural'' modality of sound? Why is speech so much more effective than other acoustic signals? Is there some peculiar, perhaps biologically ordained, relation between speech and the structure of language? We will return to these questions below.

I have not recounted these three failures of applied research missions to argue that money and effort spent on them were wasted. On the contrary, initial failure spurred researchers to revised efforts, and valuable progress has since been made. Reading machines for the blind, using an artificial speech output, have been developed and are already installed in large libraries (Cooper et al., 1984). There now exist automatic speech recognition devices that recognize vocabularies of roughly a thousand words, spoken in limited contexts by a few different speakers (Levinson and Liberman, 1981). Scientific texts with well-defined vocabularies can now be roughly translated by machine, then rendered into acceptable English by an informed human editor.

These advances have largely come about by virtue of brute computational force and technological ingenuity, rather than through real gains in our understanding of language. This is not because we have made no gains, for as we shall see shortly, we surely have. However, none of the devices that speak, listen, or understand actually speaks, listens, or understands according to known principles of human speech and language. For example, a speech synthesizer is the functional equivalent of a human speaker to the extent that it produces intelligible speech. But it obviously does so by quite different means than those that humans use: none of its inorganic components correspond to the biophysical structures of larynx, tongue, velum, lips, and jaw. Instead, a synthesizer simulates speech by means of a complex system of tuned electronic circuits, and resembles a speaker somewhat as, say, a crane resembles a human lifting a weight. We are still deeply ignorant of the physiological controls by which a speaker precisely coordinates the actions of larynx, tongue, and lips to produce even a single syllable.

In short, the main scientific value of the early work I have described was to reveal the astonishing complexity of speech and language, and the inadequacy of earlier theories to account for it. One important effect of the initial failures was therefore to prepare the ground for a theoretical revolution in linguistics (and psychology) that began to take hold in the late 1950s.

THE GENERATIVE REVOLUTION IN LINGUISTICS

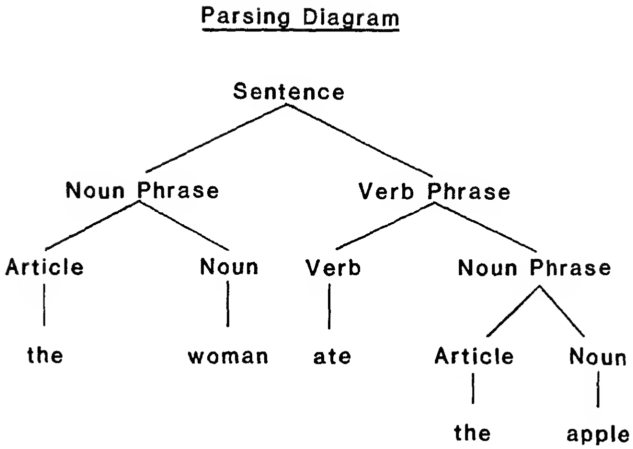
The publication in 1957 of Noam Chomsky's *Syntactic Structures* began a revolution in linguistics that has been sustained and developed by many subsequent works (e.g., Chomsky, 1965, 1972, 1975, 1980; Chomsky and Halle, 1968). To describe the course of this revolution is well beyond the scope of this chapter. However, the impact of Chomsky's writings on fields outside linguistics—philosophy, psychology, biology, for example—and their importance for the emerging science of language has been so great that some brief exposition of at least their nontechnical aspects is essential. I should emphasize that Chomsky's work has by no means gone unchallenged (e.g., Givon, 1979; Hockett, 1968; Katz, 1981). My intent in what follows is not to present a brief in its defense, but simply to sketch a bare outline of the most influential body of work in modern linguistics.

The central goal of Chomsky's work has been to formalize, with mathematical rigor and precision, the properties of a successful grammar. He defines a grammar as "a device of some sort for producing the sentences of the language under analysis" (Chomsky, 1957, p. 11). A grammar, in Chomsky's view, is not concerned either with the meaning of a sentence or with the physical structures (sounds, script, manual signs) that convey it. The grammar, or syntax, of a language is a purely formal system for arranging the words (or morphemes) of a sentence into a pattern that a native speaker would judge to be grammatically correct or at least acceptable. In *Syntactic Structures*, Chomsky compared three types of grammar: finite-state, phrase-structure, and transformational grammars.

A finite-state grammar generates sentences in a left-to-right fashion: given the first word, each successive word is a function of the immediately preceding word. (Such a model is, of course, precisely that adopted by B.F. Skinner in his *Verbal Behavior* (1957), a *dernier cri* in behaviorism, published in the same year as the ''*premier cri*'' of the new linguistics.) Chomsky (1956) proved mathematically, as work on machine translation had suggested empirically, that a simple left-to-right grammar can never suffice as the grammar of a natural language. The reason, stated nontechnically, is that there may exist dependencies between words that are not adjacent, and an indefinite number of phrases containing other nonadjacent dependencies may bracket the original pair. Thus, in the sentence, *Anyone who eats the fruit is damned, anyone* and *is damned* are interdependent. We can, in principle, continue to add bracketing interdependencies indefinitely, as in *Whoever believes that anyone who eats the fruit is damned is wrong*, and *Whoever denies that whoever believes that anyone who eats the fruit is damned is wrong is right.*

In practice, we seldom construct such sentences. However, the recursive principle that they illustrate is crucial to every language. The principle permits us to extend our communicative reach by embedding one sentence within another. For example, even a four-year-old child may combine, *We picked an apple* and *I want an apple for supper* into the utterance *I want the apple we picked for supper*. Thus, the child embeds an adjectival phrase, *we picked* (= *that we picked* with the relative pronoun deleted), to capture two related sentences in a single utterance (cf. Limber, 1973).

Chomsky goes on to consider how we might formulate an alternative and more powerful grammar, based on the traditional constituent analysis of sentences into ''parts of speech.'' Constituent analysis takes advantage of the fact that the words of any language (or an equivalent set of words and affixes) can be grouped into categories (such as noun, pronoun, verb, adjective, adverb, preposition, conjunction, article) and that only certain sequences of these categories form acceptable phrases, clauses, and sentences. By grouping grammatical categories into permissible sequences, we can arrive at what Chomsky terms a phrase-structure grammar. Such a grammar is ''a finite set . . . of initial strings and a finite set . . . of 'instruction formulas' of the form X→Y interpreted: 'rewrite X as Y' '' (Chomsky, 1957, p. 29). Figure 1 illustrates a standard parsing diagram of the utterance, *The woman ate the apple*, in a form familiar to us from grammar school (above), and as a set of ''rewrite rules'' from which the parsing diagram can be generated (below).



Rewrite Rules

(1) Sentence → Noun Phrase + Verb Phrase
(2) Noun Phrase → Article + Noun
(3) Verb Phrase → Verb + Noun Phrase
(4) Article → {the, a}
(5) Noun → {woman, apple...}
(6) Verb → {ate, seized...}

FIGURE 1 Above, a parsing diagram dividing the sentence *The woman ate the apple* into its constituents. Below, a set of rewrite rules that will generate any sentence having the constituent structure shown above.

Notice, incidentally, that rewrite rules are indifferent to meaning. They will generate anomalous utterances such as *The chocolate loved the clock*, no less readily than *The woman ate the apple*. Moreover, many native speakers would be willing to accept such anomalous utterances as grammatically correct, even though they have no meaning. This hints at the possibility that syntactic capacity might be autonomous, a relatively independent component of the language faculty. This is a matter to which we will return below.

An important point about a set of rewrite rules is that it specifies the grouping of words necessary to correct understanding of a sentence. The sentence *Let's have some good bread and wine* is ambiguous until we know whether the adjective *good* modifies only *bread* or both *bread* and *wine*. The distinction may seem trivial. But, in fact, the example shows that we are sensitive (or can be made sensitive) to an ambiguity that could not have arisen from any difference in the words themselves or in their sequence. Rather, the origin of the ambiguity lies in our uncertainty as to how the words should be grouped, that is, as to their phrase structure. A correct (or incorrect) interpretation of their meaning therefore depends on the listener (and *a fortiori* the speaker) being able to assign an abstract phrase structure to the sequence of words.

Whether a complete grammar of English, or any other natural language, could be written as a set of phrase-structure rules is not clear. In any event, Chomsky argues in *Syntactic Structures* that such a grammar would be unnecessarily repetitive and complex, since it does not capture a native speaker's intuition that certain classes of sentence are structurally related. For example, the active sentence *Eve ate the apple* and the passive sentence *The apple was eaten by Eve* could both be generated by an appropriate set of phrase-structure rules, but the rules would be different for active sentences than for their passive counterparts. Surely, the argument runs, it would be "simpler" if the grammar somehow acknowledged their structural relation by deriving both sentences from a common underlying "deep structure." The derivation would be accomplished by a series of steps or "transformations" whose functions are to delete, modify, or change the order of the base constituents *Eve, ate, apple.*

An important aspect of transformations is that they are structure dependent, that is, they depend on the analysis of a sentence into its structural components, or constitutents. For example, to transform such a declarative sentence as *The man is in the garden* into its associated interrogative *Is the man in the garden*?, a simple left-to-right rule would be: "Move the first occurrence of *is* to the front." However, the rule would not then serve for such a sentence as *The man who is tall is in the garden*, since it would yield *Is the man who tall is in the garden*? The rule must therefore be something like: "Find the first occurrence of *is* following the first noun phrase, and move it to the front" (Chomsky, 1975, pp. 30–31). Thus, a transformational grammar, no less than a phrase-structure grammar, presupposes analysis of an utterance into its grammatical (or phrasal) constituents. We may note, in passing, that children learning a language never produce sentences such as *Is the man who tall is in the garden*? Rather, their errors suggest that, even in their earliest attempts to frame a complex sentence, they draw on a capacity to recognize the structural components of an utterance.

However, here we should be cautious. Chomsky has repeatedly emphasized that ". . .a generative grammar is not a model for a speaker or hearer" (1965, p. 9), not a model of psychological processes presumed to be going on as we speak and listen. The word "generative" is perhaps misleading in this regard. Certainly, experimental psychologists during the 1960s devoted much ingenuity and effort to testing the psychological reality of transformations (for reviews, see Cairns and Cairns, 1976; Fodor et al., 1974; Foss and Hakes, 1978). But the net outcome of this work was to demonstrate the force of Chomsky's distinction between formal descriptions of a language and the strategies that speakers and listeners deploy in communicating with each other (cf. Bever, 1970).

At first glance, the distinction might seem to be precisely that between *langue* and *parole*, drawn by de Saussure. However, for de Saussure, *langue*, the system of language, "exists only by virtue of a sort of contract signed by the members of a community" (de Saussure, 1966, p. 14): it is a kind of formal artifice or convention, maintained by social processes of which individuals may be quite unaware. By contrast, for Chomsky the "generative grammar [of a language] attempts to specify what the speaker actually knows" (1965, p. 8). What a speaker knows, his *competence* in Chomsky's terminology, is attested to by "intuitive" judgments of grammaticality. What a speaker does, *performance* (*parole*), is linguistic competence filtered through the indecisions, memory lapses, false starts, stammerings, and the "thousand natural [nonlinguistic] shocks that flesh is heir to." Thus, even though a theory of grammar is not a theory of psychological process, it *is* a theory of individual linguistic capacity.

In Chomsky's view, the task of linguistics is to describe the structure of language much as an anatomist might describe the structure of the human hand. The complementary role of psychology in language research is to describe language function and its course of behavioral development in the individual, while physiology, neurology, and psychoneurology chart its underlying structures and mechanisms.

Whether this sharp distinction between language as a formal object and language as a mode of biological function can, or should, be maintained is an open question. What is clear, however, is that it was from a rigorous analysis of the formal properties of syntax (and later of phonology: see Chomsky and Halle, 1968) that Chomsky was led to view language as an autonomous system, distinct from other cognitive systems of the human mind (cf. Fodor, 1982; Pylyshyn, 1980). His writings during the late 1950s and 1960s brought an exhilarating breath of fresh air to psychologists interested in language, because they offered an escape from the stifling behavioristic impasse, already noted by Lashley (1951) and others (e.g., Miller et al., 1960).

The result was an explosion of research in the psychology of language, with a strong emphasis on its biological underpinnings. Whatever one's view of generative grammar, it is fair to say that almost every area of language study over the past 25 years has been touched, directly or indirectly, whether into action or into reaction, by Chomsky's work. This will be obvious from the following selective review of research in four major areas: acoustic phonetics, American Sign Language (ASL), brain specialization for language, and language development in children.

Acoustic Phonetics

We begin with audible speech, partly because we are then following the course of development, both in the species and the individual, from the bottom up; partly because it is in this area, where we are dealing with observable, physical processes, that the most dramatic progress has been made; and partly because we have come to realize in recent years that the physical medium of language places fundamental constraints on its surface structure. To understand this we must know something of the way speech is produced.

The Source-filter Theory of Speech Production The source-filter theory, first proposed by Johannes Müller in 1848, has been elaborated in the past 50 years, notably at the University of Tokyo (Chiba and Kajiyama, 1941), the Royal Institute of Technology in Stockholm (Fant, 1960, 1973) and, in this country, the Massachusetts Institute of Technology (Stevens and House, 1955, 1961) and Bell Telephone Laboratories (Flanagan, 1983). As a result of this work, we are now able to specify accurately the possible acoustic outputs of any vocal tract, animal or human.

When we speak, we drive air from our lungs through the pharynx, mouth, teeth, lips, and, sometimes, nose. The sound source is usually either the "voice" produced by rapid pulsing of the vocal cords (as in the final sounds of *be* and *do*), the hiss of air blown through a narrow constriction (as in the initial and final sounds of *safe* and *thrush*) or both (as in the final sounds of *leave* and *bees*). The resonant filter is the vocal tract, its air set into vibration by the flow of air from the lungs, much as we produce sound from a bottle or a wind instrument by blowing air across its top.

To some large degree linguistic information (that is, consonants and vowels) is conveyed by systematic variations in the configuration of the vocal tract. For example, if we lower the tongue and move it back toward the pharynx, we set up a pattern of resonances (known as formants) corresponding to the vowel [a]. If we raise the tongue forward toward the gums, we set up resonances for the vowel [i]. Finally, if we raise the tongue backward toward the soft palate, we set up resonances for the vowel [u]. These three sounds are the most distinct vowels, both articulatorily and acoustically, that the human vocal tract can produce, and all known languages use at least two of them.

[We may note in passing that Lieberman and his colleagues (Lieberman and Crelin, 1971; Lieberman et al., 1972) have used the source-filter theory of speech production to demonstrate that these vowels lie outside the range of sounds that could be produced either by an adult chimpanzee or by a newborn human infant. The reason for this is that the larynx in both chimpanzee and infant, is high in the throat, restricting the range of possible tongue movements. An advantage of the high larynx for the infant is that it provides an arrangement of the oral tract such that, like other mammals, the infant can suck through its mouth and breathe through its nose at the same time. Over the first six months of life, the infant's larynx lowers, a special swallowing reflex develops to prevent food entering the lungs, and the infant becomes capable of producing the vowels of the language spoken around it. The lowered larynx seems to be one of several adaptations of the vocal apparatus that have suited it for speaking as well as for eating and breathing.]

Of course, we do not speak only in vowels. Rather, we speak in runs of syllables, alternately constricting the vocal tract to form consonants, opening it to form vowels. (This repeated opening and closing of the tract produces the rises and falls of amplitude that are the basis of speech rhythm and poetic meter.) What is of interest, as we have already remarked, is that the tract configurations appropriate to particular consonants and vowels do not follow each other in linear sequence. At any instant, each articulator is executing a complex pattern of movement, of which the spatiotemporal coordinates reflect the influence of several neighboring segments. Readers may test this by slowly uttering, for example, the words *cool* and *keel*. They will find that the position of the tongue on the palate during closure for the initial consonant, [k], is slightly further back for the first word than for the second. The result of this interleaving is that, at any instant, the sound is conveying information about more than one phonetic segment, and that each phonetic segment draws information from more than one piece of sound—an obvious problem for automated speech recognition. Unfortunately, we cannot, as was at one time hoped, escape from this predicament by building a machine to recognize syllables, because similar interactions between phonetic segments occur across syllable boundaries. We see all this quite clearly if we examine a sound spectrogram.

The Sound Spectrograph The sound spectrograph was developed at Bell Telephone Laboratories during World War II, to provide a visible display of the acoustic spectrum of speech as it changes over time. Originally, it was hoped that the device would enable deaf persons to use the telephone (Potter et al., 1947), but this proved impracticable because spectrograms are formidably difficult to read (but see Cole et al., 1980).

Figure 2 is a spectrogram of the utterance *She began to read her book.* Frequency on the ordinate is plotted against time on the abscissa. Variations in relative amplitude appear as variations in the darkness of the pattern. The dark bars correspond to formants, that is, to resonant peaks in the vocal tract resonance function. Scattered patches, as at the beginning, correspond to the noise of fricatives, e.g., [f], [s], and stop consonants, e.g., [p], [b]. A series of vertical lines has been drawn, dividing the spectrogram into discrete, acoustic segments. There are 25 of these segments, even though the utterance consists of only 17 phonetic segments and 7 syllables. Some of these acoustic segments correspond more or less directly to phonetic segments: thus, segments 1 and 2 correspond to the two sounds of *she.* Segment 3, on the other hand, corresponds to the first three sounds of *began*, segments 11 and 12 to the first sound of *to*, segment 23 to the first two sounds of *book.*

The sound spectrograph revealed, for the first time, the astonishing variability of the speech signal both within and across speakers. It was also the basis for the first systematic studies of speech perception, from which we have learned which aspects of the signal carry crucial phonetic information. These studies, in turn, provided the basis for the development of speech synthesis. Thus, artificial talking machines, now being used in reading machines for the blind and in a variety of human-machine communication systems, rest squarely on the shoulders of the spectrograph.

Speech Perception Early work in speech perception was largely guided by the demands of telephonic communication. Its aim was to estimate how much distortion (by filtering, noise, peak-clipping, and so on) could be imposed on the signal without seriously reducing its intelligibility (Licklider and Miller, 1951; Miller, 1951). Two general conclusions from this work were surprising and important. First, speech is so resistant to distortion that we can throw away large parts of the signal without reducing its intelligibility. Second, intelligibility does not depend on naturalness. These two facts made it possible to learn a great deal about the important information-bearing elements in speech by stripping it down to its minimal cues.

Work of this kind was first undertaken at Haskins Laboratories in New York during the 1950s, as part of a program to develop a suitable output for a reading machine. The key research tool was the Pattern Playback, developed by F.S. Cooper (Cooper, 1950; Cooper and Borst, 1952) to reconvert the visual pattern of a spectrogram into sound. The pattern, painted on a moving acetate belt, reflects frequency-modulated light to a photocell that drives a speaker. Figure 3 illustrates an early spectrogram and its stylized copy. If the copy is passed through the playback, it produces an intelligible version of the utterance *to catch pink salmon.* The utterance

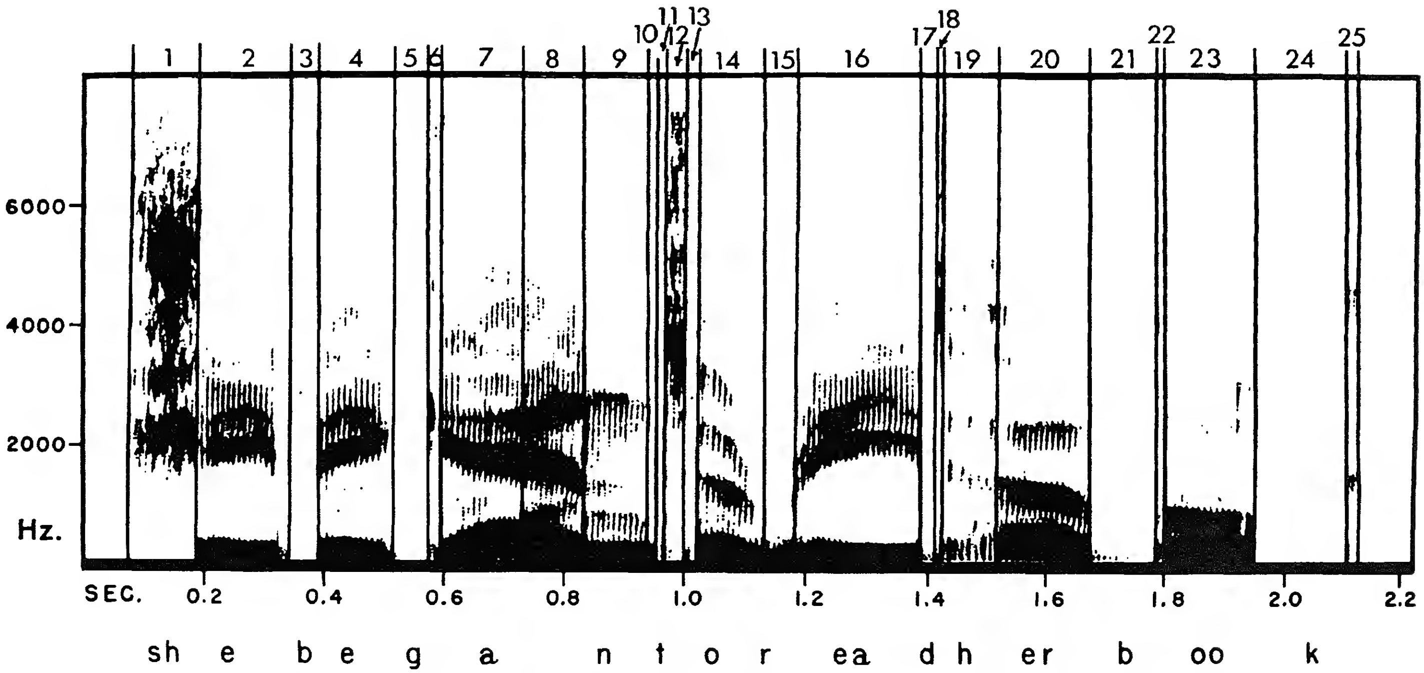


FIGURE 2 A spectrogram of the utterance *She began to read her book*. Frequency is plotted on the ordinate, time on the abscissa; relative amplitude is represented by varying degrees of darkness in the display. The dark horizontal bands reflect resonant peaks in the vocal tract transfer function (formants, conventionally numbered from the bottom up: first formant, second formant, etc.); the vertical striations reflect repeated opening and closing of the glottis (voice). Heavy vertical lines have been drawn dividing the pattern into 25 discrete acoustic segments (see text).

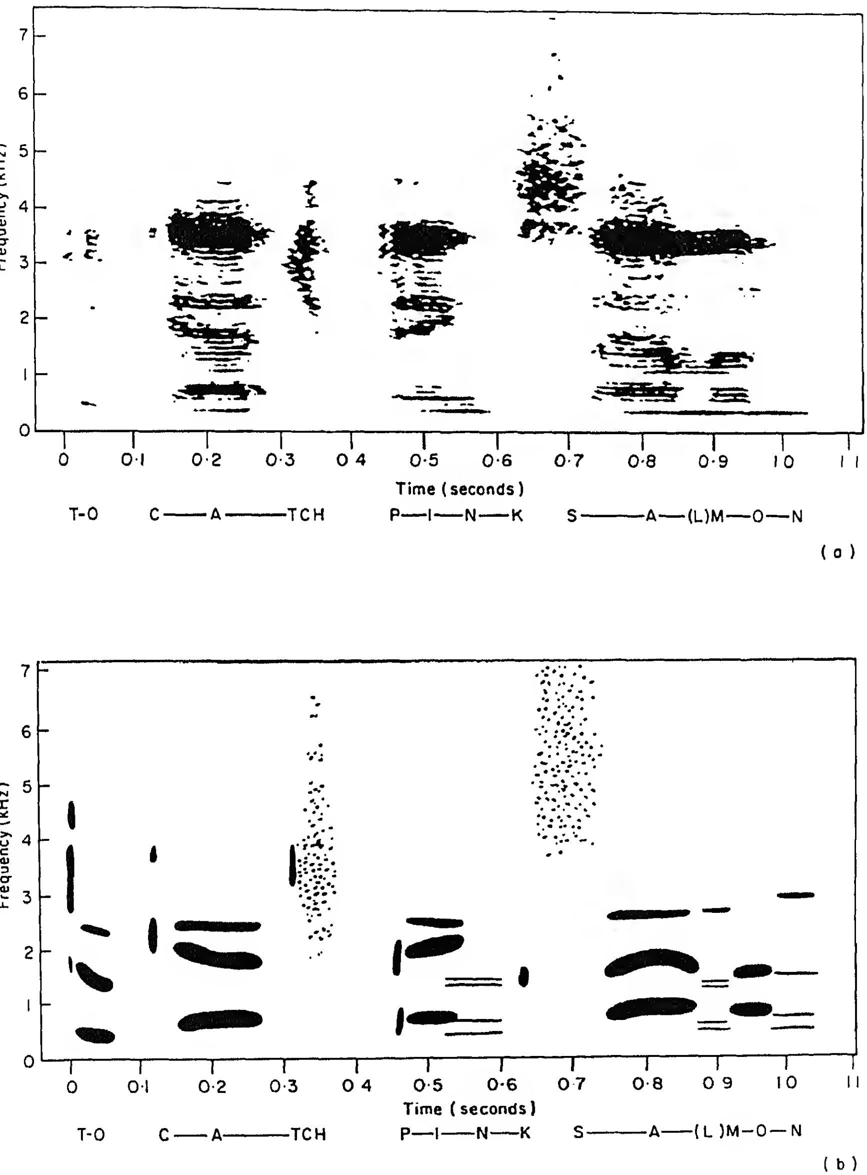


FIGURE 3 Above, a spectrogram of the utterance *To catch pink salmon*. Below, a stylized copy of the spectrogram, sufficient to regenerate the utterance if played on the Pattern Playback.

sounds unnatural, partly because the formant bandwidths have been sharply reduced, partly because it is spoken in a monotone.

The playback made it possible for experimenters to manipulate the speech signal systematically, by pruning, deleting, or exaggerating portions of the spectrographic pattern until they had determined the minimal cues for any particular utterance (Liberman, 1957; Liberman et al., 1959). With this device, and with its successors at Haskins and elsewhere, a body of knowledge was built up, sufficient for synthesis by rule of relatively high-quality speech (Fant, 1960, 1968; Flanagan, 1983; Mattingly, 1974).

Several reviews of the perceptual implications of this work have been published (Darwin, 1976; Liberman et al., 1967; Liberman and Studdert-Kennedy, 1978; Studdert-Kennedy, 1974, 1976), and I will not review them here. However, two facts deserve note. First, the cues for a given phonetic segment (that is, for a particular consonant or vowel) vary markedly as a function of context. Figure 4 displays spectrograms of the naturally spoken syllables [did] and [dud]. We know from synthetic speech that a main cue to the initial [d] lies in changes in the second formant after onset. Notice that the second formant rises before [i], falls before [u], and that the rising and falling patterns are precisely reversed for the final [d]. Yet all are heard as [d]. Moreover, if these patterns or their synthetic versions are removed from context and presented to listeners for judgments, they are no longer heard as

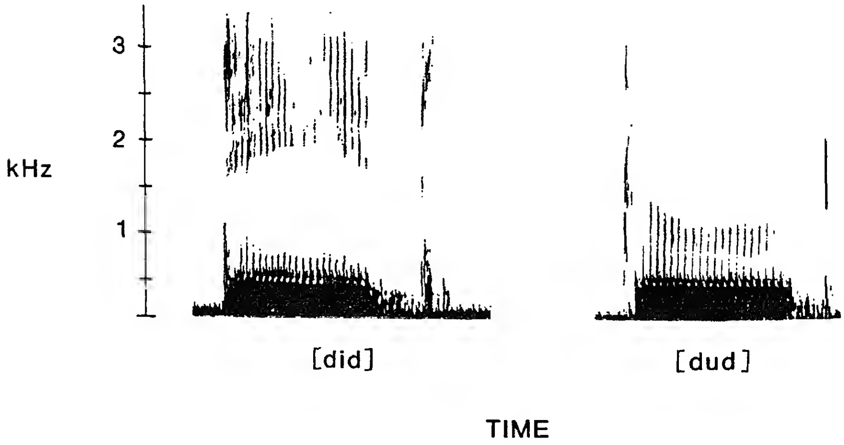


FIGURE 4 Spectrograms of naturally spoken [did] (*deed*) and [dud] (*dood*). The acoustic information specifying the alveolar place of articulation of the initial and final consonants is primarily carried by the second formant, centered around 2 kHz for [did] and slightly below 1 kHz for [dud]. Note that this formant forms a parabola, concave downwards in [did], concave upwards in [dud]. Despite this difference, both patterns are heard as beginning and ending with [d].

[d], nor are they heard as invariant. Rather they are heard as rising and falling tones (Liberman et al., 1967). In other words, different acoustic patterns are heard as different in a nonspeech context but as the same in a speech context. This is merely one of dozens of such examples.

The second fact of note is that despite the apparent lack of discrete phonetic segments in the signal, listeners have little difficulty in learning to find segments—so little, in fact, that a segmental representation of speech is the basis of the alphabet.

The interpretation of these facts is still a matter of controversy (e.g., Cole and Scott, 1974; Ladefoged, 1980; Stevens, 1975), and I will not pursue the matter here. However, it is worth noting that such findings gave rise to the hypothesis that humans have evolved a specialized perceptual mechanism for speech, distinct from, though dependent on, their general auditory system (Liberman, 1970, 1982; Liberman and Studdert-Kennedy, 1978; Liberman et al., 1967). The hypothesis has received substantial support from many dozens of studies of dichotic listening over the past 20 years (e.g., Kimura, 1961, 1967; Shankweiler and Studdert-Kennedy, 1967; Studdert-Kennedy and Shankweiler, 1970; for a review, see Porter and Hughes, 1983). The conclusion from this work, and from studies of patients with separated cerebral hemispheres (see section below on brain specialization for language), is that the left hemisphere of most normal right-handed individuals is specialized not only for speaking (as has been known for many years from studies of brain-damaged patients), but also for perceiving speech. Specifically, there is now good reason to believe that "while the general auditory system common to both hemispheres is equipped to extract the auditory parameters of a speech signal, the dominant [i.e., left] hemisphere may be specialized for the extraction of linguistic features from these parameters" (Studdert-Kennedy and Shankweiler, 1970, p. 579).

An important implication of this conclusion is that speech forms an integral part of the left-hemisphere language system discussed below. With this in mind let us turn to recent work on American Sign Language, which draws on a different perceptuomotor system from that of spoken language.

AMERICAN SIGN LANGUAGE

Speech is the natural medium of language. Specialized structures and functions have evolved for spoken communication: vocal tract morphology; lip, jaw, and tongue innervation; mechanisms of breath control (Lenneberg, 1967); and perhaps even (as I have just suggested) matching perceptual mechanisms. But is there any further specialization for language? Is language an autonomous system, distinct from other cognitive systems, as Chomsky has argued?

An opportunity to address this question has arisen in recent years from an unexpected quarter: sign languages of the deaf. Until some 20 years ago, it was commonly believed that sign languages of the deaf—and of other social groups, such as American Plains Indians and Australian aborigines—were either more or less impoverished hybrids of conventional iconic gesture and impromptu pantomime, or artificial systems based, like reading and writing, on a specific spoken language. Artificial systems, such as Signed English and Paget-Gorman, are indeed used in many schools of the deaf: their signs refer to letters (finger-spelling) or higher-order linguistic units (words, morphemes), and their syntax follows that of the base language. However, there are other signed languages, not based on any spoken language, with their own independent lexicons and syntactic systems. The most extensively studied of these is American Sign Language (ASL), the first language of over 100,000 deaf individuals and, according to Mayberry (1978), the fourth most common language (after English, Spanish, and Italian) in the United States.

Modern ASL stems from a French-based sign language introduced into the United States by Thomas Gallaudet in 1817. (According to Stokoe [1974] ASL signers today find French SL more intelligible than British SL, a nice demonstration that ASL is independent of English.) Thus, the original language was in fact based on a spoken language. However, over the past 165 years it has developed among the deaf into an independent sign language.

Structural analysis of ASL was first undertaken by Stokoe (1960), and in 1965 he and his colleagues, Casterline and Croneberg (Stokoe et al. 1965), published *A Dictionary of American Sign Language on Linguistic Principles*, containing a description and English gloss of nearly 2,500 signs. The dictionary used minimal pair analysis to show that signs contrasted along three independent dimensions: hand configuration, place of articulation, and movement. For example, signs for APPLE and JEALOUS contrast in hand configuration; signs for SUMMER and UGLY contrast in place of articulation; signs for CHAIR and TRAIN contrast in movement (Klima and Bellugi, 1979, p. 42). Stokoe and his colleagues isolated 55 "cheremes" or primes, analogous to the phonemes of a spoken language: 19 for hand configuration, 12 for place of articulation, and 24 for movement. Thus, they demonstrated that ASL has a sublexical structure, analogous to the phonological structure of a spoken language.

ASL also has a second level of structure, a grammar or syntax. This has been demonstrated in an extensive program of research at the Salk Institute for Biological Studies in La Jolla, California, over the past 10 years (Klima and Bellugi, 1979). I will not attempt to review this work in any detail, but several points deserve note. First, ASL has a rule-governed system of compounding, by which signs may be combined to form a new sign different in meaning from its components. The process is analogous to that by which, in English, *hard* and *hat*, say, are combined to form *hardhat*, meaning a construction worker. Thus, the lexicon of ASL can be expanded by rule, not simply by iconic invention.

Second, ASL has an elaborate system of inflections by which it modulates the meaning of a word. For example, in English, changes in aspectual meaning (that is, distinctions in the onset, duration, frequency, recurrence, permanence, or intensity of an event) are indicated by concatenating morphemes. We may say, *he is quiet, he became quiet, he used to be quiet, he tends to be quiet*, and so on. All these meanings are conveyed in ASL by distinct modulations of the root sign's movement. In the root sign for QUIET the hands move straight down from the mouth, while for TENDS TO BE QUIET they move down forming a circle. Similarly, related nouns and verbs are also distinguished by movements, while verbs are inflected by movement modulation for person, number, reciprocal action, and aspect.

Third, ASL has a spatial (rather than a temporal) syntax. Nouns introduced into a discourse are assigned arbitrary reference points in a horizontal plane in front of the signer. These points then serve to index grammatical relations among referents: verb signs are executed with a movement between two points, or across several points, to indicate subject and object. Thus, a grammatical function variously served in spoken language by word order, case markers, verb inflections, and pronouns is fulfilled in ASL by a spatial device.

Finally, ASL has a variety of syntactic devices that make use of the face. Liddell (1978) has shown that a relative clause ("The apple *that Eve offered* tempted him") may be marked by tilting back the head, raising the eyebrows, and tensing the upper lip for the duration of the clause. Baker and Padden (1978) describe gestures of the face and head that mark the juncture of conditional clauses ("*If you eat the fruit*, you will be punished").

In short, though structural analysis of ASL is far from complete, it is evident that the language has a dual pattern of form and syntax, fully analogous to that of a spoken language. Nonetheless, there are differences. The main structural difference between ASL and English was illustrated by Klima and Bellugi (1979) in a comparison of their rates of communication. The times taken to tell a story in the two languages were almost exactly equal. Yet the speaker used two to three times as many words as the signer used signs. The reason for the discrepancy, already hinted at, lies in the temporal distribution of information. Speech, for the most part, develops its patterns in time, sequentially, while ASL develops its patterns both simultaneously, in space, and sequentially. The difference is evidently due to the difference in the perceptual modalities addressed. Sign, addressed to the eye, is free to package information in parallel; speech, addressed to the ear, is forced into a serial mode. What is interesting, of course, is that despite constraints of modality, the two languages convey information at roughly the same rate. This suggests that they may be operating under the same temporal constraints of cognition.

What, finally, are the implications of this work for the study of speech and language? Evidently, the dual structure of language is not a mere consequence of perceptuomotor modality but a reflection of cognitive requirements. Whether these cognitive requirements are linguistic rather than general is still not clear. Differently put, we still do not know whether the relation between signed and spoken language is one of analogy or homology. If the two systems prove to be homologous, that is, if they prove to draw on the same neural structures and organization, we will have strong evidence that language is a distinct cognitive faculty. However, if they do not draw on the same underlying neural organization, we might suppose that linguistic structure is purely functional, the adventitious consequence of a cognitively complex animal's attempt to communicate its thought. Studies of sign-language breakdown due to brain injury, discussed below, are therefore of unusual interest and importance.

BRAIN SPECIALIZATION FOR LANGUAGE

Most of our knowledge of brain specialization for language comes from those ''experiments of nature'' in which some more or less circumscribed lesion (due to stroke, epilepsy, congenital malformation, gunshot wounds, and so on) proves to be correlated with some more or less circumscribed cognitive or linguistic deficit (for a brief account of modern brain-scanning techniques, see Benson, 1983, and references therein). Recently, our sources of knowledge have been expanded by use of brain stimulation, preparatory to surgery under local anesthesia (Ojemann, 1983, and references therein), and by studies of so-called ''split-brain'' patients whose cerebral hemispheres have been separated surgically for relief of epilepsy (see below). Some degree of concordance between patterns of brain localization in normal and abnormal individuals has been established by experiments on normals in which visual or auditory input is confined, or more clearly delivered, to one hemisphere rather than the other (Moscovitch, 1983).

Evidence From Studies of Aphasia

The term *aphasia* refers to some impairment in language function, whether of comprehension, production, or both, due to some more or less well localized damage to the brain. Systematic study of aphasia goes back well over a hundred years, and the literature of the subject is vast (for reviews, see, for example, Goodglass and Geschwind, 1976; Hecaen and Albert, 1978; Lesser, 1978; Luria, 1966, 1970). The most that can be done here is to hint at one area in which linguistics (that is, formal language description) has begun to affect aphasia studies.

Until recently, the standard framework for describing aphasic symptoms was that of the language modalities: speaking, listening, reading, and writing, or, more generally, the dimensions of expression and reception. These are still the dimensions of the major test batteries used to diagnose aphasia, such as the Boston Diagnostic Aphasia Examination (Goodglass and Kaplan, 1972). An important assumption, underlying any attempt at diagnosis, is that damage to a particular region of the brain has particular, not general, effects on language function. The assumption has strong empirical support and has led to the isolation of two (among several other) broad types of aphasia, nonfluent and fluent, respectively associated with damage to the left cerebral hemisphere in an anterior region around the third frontal convolution (Broca's area) and a posterior region around the superior temporal convolution (Wernicke's area).

Broca's area lies close to the motor strip of the cortex (in fact, close to that portion of the strip associated with motor control of the jaw, lips, and tongue), while Wernicke's area surrounds the primary auditory region. In accord with this anatomical dissociation, a Broca's aphasic (that is, an individual with damage to Broca's area) has been classically found to be nonfluent: having good comprehension but awkward speech, characterized by pauses, difficulties in word-finding and distorted articulation; utterances are described as ''telegrammatic,'' consisting of simple, declarative sentences, relying on nouns and uninflected verbs, omitting grammatical morphemes or function words. By contrast, a Wernicke's aphasic has been found to have poor comprehension, even of single words, but fluent speech, composed of inappropriate or nonexistent (though phonologically correct) words, often inappropriately inflected and/or out of order.

Notice that these descriptions are still couched in terms of input and output—that is, modalities of behavior—rather than in linguistic terms. The idea that linguistic theory should be brought to bear on aphasia, and attempts made to characterize deficits in terms of overarching linguistic function, has been proposed a number of times in the past (e.g., Jakobson, 1941; Pick, 1913). But only recently (again, partly under the influence of Chomsky's view of language as an autonomous system, composed of autonomous syntactic and phonological subsystems) has the idea begun to receive widespread attention. The general hypothesis of the studies described below is that language breaks down along linguistic rather than modal lines of demarcation.

We will focus mainly on the hypothesis that syntactic com discretely and coherently represented in Broca's area of the left fr If this is so, the clinical impression that Broca's aphasics have g prehension, despite their agrammatic speech (and, incidentally must be in error. More careful testing should reveal deficits in prehension, also.

Caramazza and Zurif (1976) tested this hypothesis with thre sentences: (1) simple declarative sentences in which semantic might permit decoding without appeal to syntax (*The apple that eating is red*); (2) so-called reversible sentences that require kno syntactic relations for decoding (*The boy that the girl is chasi* and (3) implausible, though grammatically correct, sentences (*Th the dog is patting is fat*). The sentences were presented orally, a were asked to choose which of two pictures represented the mea sentence. The incorrect alternative showed either a subject-obje or an action different from that specified by the verb.

Broca's aphasics performed very well on simple declarative and on sentences with strong semantic constraints (as when th alternative depicted the wrong action). On reversible plausibl plausible sentences (when the incorrect alternative depicted a sub reversal) the patients' performance was at chance. Caramazza (1976) concluded that the clinical impression of good compre Broca's aphasics was due to their ability to draw on semantic and constraints to understand sentences despite their inability to proc

Other studies have shown that Broca's aphasics have difficulty a sentence into its grammatical constituents (Von Stockert, 19 use articles to assign appropriate reference in understanding (Goodenough et al., 1977); and cannot, in general, access c grammatical morphemes (Zurif and Blumstein, 1978). These not without their critics (e.g., Linebarger et al., 1983), nor is claim that aphasic breakdown is typically (or, indeed, ever) al linguistic lines (Studdert-Kennedy, 1983, pp. 193–194): the loc tent of brain damage in aphasia is largely a matter of chance, a that language alone is affected. However, we have other sources to test the hypothesis that syntax is represented in the brain as a f discrete subsystem.

Evidence From Split-brain Studies

One source of evidence is the split-brain patient whose cere spheres have been separated surgically for relief of epilepsy. Th permits an investigator to assess the cognitive and linguistic ca each hemisphere separately. Zaidel (1978) has devised a contact lens, opaque on either the nasal or temporal side, that can be used (profiting from decussation of the optic pathways) to ensure that visual information is freely scanned by a single hemisphere. A variety of written verbal materials—nonsense syllables, words, sentences of varying length and complexity—and pictures can then be used to test the capacities of the isolated hemispheres. For example, the sentences, *The fish is eating* or *The fish are eating*, can be presented to a single hemisphere, together with appropriate alternative pictures, to test the hemisphere's capacity to understand written verbal auxiliaries (*is, are*) (Zaidel, 1983). Similarly, pictures of various objects belonging to different classes (fruit, furniture, vehicles, etc.) might be presented to a single hemisphere to test the hemisphere's capacity to categorize.

The number of available subjects is, of course, limited. But the conclusions from studies of four split-brain patients are remarkably consistent (Zaidel, 1978, 1980, 1983). In general, each hemisphere seems to have "a complete cognitive system with its own perception, memory, language, and cognitive abilities, but with a unique profile of competencies: good on some abilities, poor on others" (Zaidel, 1980, p. 318). Of particular interest in the present context is the finding that, although the right hemisphere cannot speak, it has a sizable auditory and reading lexicon. However, unlike the left hemisphere, the right cannot read new (nonsense or unknown) words or recognize words for which it has no semantic interpretation. Similarly, the right hemisphere cannot group pictures of objects on the basis of rhyme (e.g., *nail, male*). Evidently, phonological analysis is the prerogative of the left hemisphere.

The syntactic capacity of the right hemisphere is also limited. The hemisphere can recognize verbal auxiliaries (see above), but has difficulty in discriminating inflections (*The fish eat* versus *The fish eats*). Similarly, the right hemisphere can recognize and interpret nouns, adjectives, and certain prepositions, but has difficulty with the English infinitive marker *to*. These findings on closed-class morphemes mesh to a degree with the deficits of Broca's aphasics, described above. Not surprisingly, the right hemisphere's capacity to understand sentences is sharply reduced: it cannot deal with sentences longer than about three words.

On the evidence of these studies, then, the right hemisphere has essentially no phonological capacity and only a limited syntactic capacity. Unfortunately, the limited syntactic capacity is equivocal because all these split-brain patients have had epilepsy since early childhood. Brain disorders are known to lead to reorganization and redistribution of function, particularly in childhood (Lenneberg, 1967; Dennis, 1983). We cannot therefore be sure that such syntactic capacity as the right hemisphere displays does not reflect compensation for left hemisphere deficiencies, induced by epilepsy.

Evidence From Studies of ASL "Aphasia"

Studies of normally hearing, brain-damaged patients have established a double dissociation of brain locus and function in right-handed individuals: the left cerebral hemisphere is specialized for language, the right hemisphere for visual-spatial functions (as revealed, for example, by tests requiring a subject to copy a drawing, assemble wooden blocks into a pattern, or discriminate between photographs of unfamiliar faces). As we have seen, ASL is an autonomous linguistic system with a dual structure analogous to that of spoken language, on the one hand; yet, on the other, it encodes its meanings in visual-spatial rather than auditory-temporal patterns. How then should we expect brain damage to affect the language of a native ASL signer?

The answer bears directly on our understanding of the basis of brain specialization for language. For if language loss in ASL aphasia follows damage to the right hemisphere, we may infer that language is drawn to the hemisphere controlling its perceptuomotor channel of communication. But if language loss follows damage to the left hemisphere, we may infer that the neural structure of that hemisphere is, in some sense, matched to the structure of language, whatever its modality. Language might then be seen as a distinct cognitive faculty, sufficiently abstract in its descriptive predicates to encompass both speaking and signing.

Recent studies at the Salk Institute, the first systematic and linguistically motivated studies of ASL aphasia on record, support the second hypothesis. Moreover, the forms of ASL breakdown vary with locus of lesion in a fashion strikingly similar to certain forms of spoken-language breakdown. Bellugi, Poizner, and Klima (1983) describe three patients, all of whom are native ASL signers and display normal visual-spatial capacity for non-language functions. Their symptoms, resulting from strokes, divide readily into the two broad classes noted above for spoken language: two patients are fluent, one is nonfluent.

The two fluent patients display quite different symptoms, coordinated with different areas of damage to the left hemisphere. The deficits of one patient (PD) are primarily grammatical; the deficits of the other (KL) are primarily lexical. PD has extensive subcortical damage from below Broca's area in the frontal lobe through the parietal to the temporal lobe, abutting Wernicke's area. PD produces basically normal root signs, but displays an abundance of semantic and grammatical paraphasias. He produces many semantically displaced signs (e.g., EARTH for ROOM, BED for CHAIR, DAUGHTER for WIFE). More strikingly, he often modulates an appropriate root form with an inappropriate or nonsensical inflection. Finally (despite his normal nonlanguage visual-spatial capacity), his spatial syntax is severely disordered: he misuses or avoids spatial indexing (the equivalent of pronominal function, as noted above), and overuses nouns.

The second fluent patient, KL, has more limited damage, extending in a strip across the left parietal lobe. Her deficits, though relatively mild, are almost the reverse of PD's. First, she avoids nouns and overuses pronouns (spatial indexing). Second, she tends to make formational errors in root signs, producing nonsense items by substituting incorrect hand configurations, places of articulation, or movements. Thus, these two fluent patients display almost complementary deficits, breaking along linguistic fault lines, as it were, between lexicon and grammar.

The third patient (GD) is nonfluent. She has massive damage over most of the left frontal lobe, including Broca's area. She produces individual signs correctly (with her nondominant hand, due to paralysis of the right side of her body), and can repeat a test series of signs rapidly and accurately, so that her deficits are not simply motoric. Yet her spontaneous signing invites description by just those epithets that characterize a Broca's aphasic. Her utterances are slow, effortful, short and agrammatic, largely made up of open-class items. She omits all grammatical formatives, including inflections, morphological modulations, and most spatial indices. In short, this patient, too, displays a peculiarly linguistic rather than a general cognitive pattern of breakdown.

From this brief review of brain specialization for language we may draw several conclusions. First, language breakdown seems to follow rough linguistic lines of demarcation, indicating that phonology (or patterns of sign formation) and syntax may be supported by separable neural subsystems within the left hemisphere. Second, left hemisphere specialization does not rest on a particular sensorimotor channel. Rather, the hemisphere supports general linguistic functions, common to both spoken and signed language. Thus, despite the left hemisphere's innate predisposition for speech (see section below on language acquisition), its initial neural organization is sufficiently plastic to admit quite different language forms (cf. Neville, 1980; Neville et al., 1982). At the same time, we still do not know enough about the anatomy and physiology of the brain to be sure that areas important for particular functions in spoken language precisely correspond to areas important for analogous functions in signed language: the issue of analogy versus homology is not yet closed.

Several further cautions should be noted. It is not yet clear (either from linguistic theory or from behavioral evidence) that syntax and phonology constitute homogeneous functions: some aspects of syntax and phonology may be separable from some aspects of language, others may not (Dennis, 1983). Second, it is even less clear that we should expect a coherent function, once specified, to be discretely and coherently localized in the brain. In looking for correspondences between one level of description (linguistic) and another level (neurological), we may be guilty of the "first-order isomorphism fallacy" that caused the downfall of phrenology and faculty psychology. The error would be analogous to that of someone who expected a single function of an automobile—say, acceleration—to be discretely and coherently localized in the engine. In fact, of course, the mechanism underlying acceleration is distributed over gears, fuel pump, carburator, pistons, and so on. Perhaps syntactic and phonological functions emerge, like acceleration, from the coordinated actions of disparate parts.

LANGUAGE ACQUISITION

As many as 5 percent of American children suffer from some form of delayed or disordered language development, and many more join the ranks of the illiterate. Moreover, there is growing evidence that the capacity to read depends in large part on normal development of the primary language processes of speaking and listening (Crain and Shankweiler, in press). Scientific understanding of development is therefore of broad pediatric and educational interest. In the first instance, the work may simply permit us to establish reliable norms, based on a sound understanding of what language acquisition entails. Later, we may hope, the work should lead to more effective therapeutic intervention than is now available.

No area of language study has been more strongly affected by Chomsky's work than language acquisition. Indeed, it is fair to say that until Chomsky's writings began to be widely disseminated among psychologists, in the early 1960s, the field did not exist. The few psychologists who considered the matter at all (e.g., Mowrer, 1960; Skinner, 1957) assumed that language learning would be subsumed under the general learning theory that behaviorists were striving to develop. Yet today the field has grown to such depth and complexity that a recent volume on the state of the art (Wanner and Gleitman, 1982) lists some 900 references, over half of them published in the last 10 years. The most that I can hope to do here is sketch some of the reasons for this phenomenal growth. What did Chomsky say that aroused such interest? What questions are researchers trying to answer?

Language development is a central issue in Chomsky's thought (e.g., 1965, 1972, 1980), bearing directly on the natural categories of the human mind. The issue arises from four assumptions. First, any grammar sufficient to generate the sentences of a natural language is a complex "system of many . . . rules of . . . different types organized in accordance with certain fixed principles of ordering and applicability and containing a certain fixed substructure'' (1972, p. 75). Second, the descriptive predicates of this system (grammatical categories, phonological classes) are not commensurate with those of any other known system in the world or in the mind. Third, the data available to the child in the speech of others is ''meager and degenerate.'' Fourth, no known theory of learning—least of all a stimulus-response reinforcement theory of the kind scathingly criticized by Chomsky in his review (1959) of Skinner's *Verbal Behavior* (1957)—is adequate to account for a child's learning a language. Chomsky (1972) therefore assigns to the mind an innate property, a schema constituting the ''universal grammar'' to which every language must conform. The schema is highly restrictive, so that the child's search for the grammar of the language it is learning will not be impossibly long.

Chomsky (1972) then divides the research task into three parts. First is the linguist's task: to define the essential properties of human language, the schema or universal grammar. Second is the psychologist's task of determining the minimal conditions that will trigger the child's innate linguistic mechanisms. The third task, closely related to the second, arises from the assumption that most of the utterances a child hears are not well formed. How then is the child to know which utterances to accept as evidence of the grammar it is searching for and which utterances to reject? The third task is therefore to discover the nature of the relation between a set of data and a potential grammar, sufficient to validate the grammar as a theory of the language being learned.

The proposition that language is an innate faculty of the human mind has a long history in Western thought from Plato to Darwin. The proposition is logically independent of any particular theory of language structure. Indeed, the entire enterprise of generative grammar might fail, yet leave the claim of innateness untouched. Certainly Chomsky's linguistic theories have been, and continue to be, a rich source of hypothesis and experiment in studies of language acquisition. However, his principle achievement in this area has been to force recognition that the learning of a language is an extraordinarily complex process with profound implications for the nature of mind. He has formulated the problem of language learning more precisely than ever before, spelling out its logical prerequisites in a fashion that promises to lead, given appropriate research, to a more precise specification of the innate ''knowledge'' that a child must bring to bear if it is ever to learn a language at all.

As we have noted, Chomsky's challenge precipitated a vast quantity of research. The first need was for data, for systematic descriptions of how language actually develops. Work initially concentrated on syntactic development (e.g., Brown, 1973), but in the past dozen years has expanded to include phonology (e.g., Yeni-Komshian et al., 1980), semantics (e.g., Carey, 1982; MacNamara, 1982), and pragmatics (e.g., Bates and MacWhinney, 1982). As data have accumulated it has become possible to answer many questions and, of course, to ask many more.

When does language development begin? Can we isolate reliable stages of development across children? Do the same stages occur in different language environments? Is the input to the child truly "meager and degenerate"? Is the child really constructing a grammar? Is the process passive, or must the child actively engage itself? What is the role of imitation? Do we have to posit innate proclivities? If so, are they indeed purely linguistic? And so on.

To see the force of these questions, we must have a sense of the complexity of the task that faces a child learning its native language. From our discussion of the problems of speech perception and automatic speech recognition, it will be obvious that we have much to learn about how the infant discovers invariant phonetic and lexical segments in the speech signal. We still do not know how the infant learns the basic sound pattern of a language during its first two years of life and comes to speak its first few dozen words. But let us set these puzzles aside and go straight to early syntax, where the bulk of child language research has been concentrated. The goal of this work has been to infer from a child's utterances (*performance*) what it "knows" (*competence*) about grammar and the meanings encoded by grammar, at each stage of its development.

Consider, as an example, the sentence cited above, *I want the apple we picked for supper*, a sentence comfortably within the competence of a four-year-old child. What must a child know to produce such a sentence? We will look at three aspects of its structure to illustrate the basis of Chomsky's claim that grammatical categories do not map in any simple way onto the categories of general cognition.

(1) *Word order* A child who utters the sentence evidently knows the standard subject-verb-object (SVO) order of English and so says, *I want the apple*. The child does not say as (transposing into English) a Turkish or Japanese child might say, *I the apple want* (SOV) or *The apple I want* (OSV). Presumably, the English-speaking child has long since learned that *Adam loves Eve* does not mean the same as *Eve loves Adam*. A Turkish or Japanese child, on the other hand, would have learned that uncertainties, due to variable word order, as to the underlying relations expressed in a sentence (who does what to whom) are resolved by attaching appropriate suffixes to subject and object (Slobin, 1982).

So far, the mapping between grammar and world, in the three languages, would seem to be arbitrary but direct. However, we are given pause by another phrase in our example, *the apple we picked* (= *the apple that we picked*). Here, in an object relative clause, the order of subject (*we*) and object (*apple*) is reversed, and the verb (*picked*) appears at the end, giving OSV. The switch from SVO (*we picked that*) to OSV (*that we picked*) is obligatory in English object relative clauses. Notice that, to apply this rule, a child cannot draw on any knowledge of the world; rather, it must (in some sense) know the grammatical structure of the sentence. We have here, then, another example of structure dependence, noted above in our discussion of interrogatives.

(2) *Use of the article* The child says, *I want the apple*, not *I want an apple*. Of course, if many apples had been picked, *an apple* would have been correct. The distinction between definite and indefinite articles seems natural to an English speaker. To a speaker of Russian, Chinese, or other languages in which articles are not used, the distinction might seem tiresome and unnecessary. In fact, rules for use of articles in English are complex and, with respect to the aspects of the world that they encode, seemingly arbitrary. Yet the rules are learned by the third or fourth year of life (Brown, 1973, p. 271).

(3) *Noun phrases* As a final example, consider the noun phrase *the apple we picked*. These four words (article + noun + adjectival phrase) form the grammatical object of the sentence. A child who utters them must already know the general rule for constructing noun phrases in English: the adjective goes before the noun (*the red apple*), not, as in French, after the noun (*la pomme rouge*). However, there is an exception to the rule: if the adjective is itself a phrase (that is, a relative clause: *that we picked*), the adjective must follow the noun (*the apple we picked*, not *the we picked apple*). Once again, the child reveals in its utterance knowledge of a rule of English grammar that cannot be derived from knowledge of the world.

In short, there are solid grounds for believing that language structure (both at the level of sound pattern, or phonology, and at the level of syntax) may be *sui generis*. With this in mind let us briefly review some of what we know about the course of development, with particular attention to the questions with which we began.

The infant is biologically prepared to distinguish speech from nonspeech at, or very soon after, birth. A double dissociation of the left cerebral hemisphere for perceiving speech and of the right hemisphere for perceiving nonspeech sounds within days of birth has been demonstrated both electrophysiologically (e.g., Molfese, 1977) and behaviorally (e.g., Segalowitz and Chapman, 1980). Further, dozens of experiments in the past 10 years have shown that infants, in their first six months of life, can discriminate virtually any adult speech contrast from any language on which they are tested (e.g., [b] versus [p], [d] versus [g], [m] versus [n]) (Aslin et al., 1983; Eimas, 1982). There is also evidence that infants begin to recognize the function of such contrasts, to distinguish words in the surrounding language, during the second half of their first year (Werker, 1982). (For fuller review, see Studdert-Kennedy, 1985.)

In terms of sound production, Oller (1980) has described a regular progression from simple phonation (0–1 months) through canonical babbling (7–10 months) to so-called variegated babbling (11–12 months). The phonetic inventory of babbled sounds is strikingly similar across many languages and even across hearing and deaf infants up to the end of the first year (Locke, 1983). These similarities argue for a universal, rather than language-specific, course of articulatory development.

However, around the end of the twelfth month, when the child produces its first words, the influence of the surrounding language becomes evident. From this point on, universals become increasingly difficult to discern, because whatever universals there may be are masked by surface diversity among languages. In this respect, the development of language differs from the development of, say, sensorimotor intelligence or mathematical ability (cf. Gelman and Brown, this volume). Nonetheless, we can already trace some regularities across children within a language and, to some lesser extent, across languages.

The most heavily studied stage of early syntactic development, in both English and some half-dozen other languages, is the so-called two-morpheme stage. Brown (1973) divides early development into five stages on the basis of mean length of utterance (MLU), measured in terms of the number of morphemes in an utterance. The stages are "not . . . true stages in Piaget's sense" (Brown, 1973, p. 58), but convenient, roughly equidistant points from MLU = 2.00 through MLU = 4.00. The measure provides an index of language development independent of a child's chronological age.

Of interest in the present context is that no purely grammatical description of Stage I (MLU = 2.00, with an upper bound of 5.00) has been found satisfactory. Instead, the data are best described by a "rich interpretation," assigning a meaning or function to an utterance on the basis of the context in which it occurs. Brown lists eleven meanings for Stage I constructions, including: naming, recurrence (*more cup*), nonexistence (*all gone egg*), agent and action (*Mommy go*), agent and object (*Daddy key*), action and location (*sit chair*), entity and location (*Baby table*), possessor and possession (*Daddy chair*), entity and attribute (*yellow block*). Brown (1973) proposes that these meanings "derive from sensorimotor intelligence, in Piaget's sense . . . [and] probably are universal in humankind but not . . . innate" (p. 201).

We should emphasize that these Stage I patterns reflect semantic, not grammatical, relations even though they may be necessary precursors to the grammatical relations that develop during Stage II (MLU = 2.50, with an upper bound of 7.00). Brown (1973) traced the emergence of 14 grammatical morphemes in three Stage II English-speaking children. The morphemes included: prepositions (*in, on*), present progressive (*I am playing*), past regular (*jumped*), past irregular (*broke*), plural -s, possessive -s, third persons -s (*he jumps*), and others. The remarkable finding was that all three children acquired the morphemes in roughly the same order (with rank order correlations between pairs of children of 0.86 or more). This result was confirmed in a study of 21 English-speaking children by de Villiers and de Villiers (1973).

However, unlike the meanings and functions of Stage I, the more or less invariant order of morpheme acquisition of Stage II has not been confirmed for languages other than English. Perhaps we should not expect that it will be. Languages differ, as we have seen, in the grammatical devices that they use to mark relations within a sentence. The devices used by one language to express a particular grammatical relation may be, in some uncertain sense, "easier" to learn than the devices used by another language for the same grammatical relation. Slobin (1982) has compared the ages at which four equivalent grammatical constructions are learned in Turkish, Italian, Serbo-Croatian, and English. In each case, the Turkish children developed more rapidly than the other children. If these results are valid and not mere sampling error, the "studies suggest that Turkish is close to an ideal language for early acquisition" (Slobin,1982, p. 145).

Unless we suppose that Turkish parents are more attentive to their children's language than Italian, Serbo-Croatian, and English parents, we may take this result as further evidence that "selection pressures" (reinforcement) have little role to play in language learning. Brown and Hanlon (1970) showed some years ago that parents tend to correct the pronunciation and truth value, rather than the syntax, of their children's speech. Indeed, one of the puzzles of language development is why children improve at all. At each stage, the child's speech seems sufficient to satisfy its needs. Neither reinforcement nor imitation of adult speech suffices to explain the improvement. Early speech is replete with forms that the child has presumably never heard: *two sheeps, we goed, mine boot*. These errors reflect not imitation, but over-generalization of rules for forming plurals, past tenses, and possessive adjectives.

We come then to a guiding assumption of much current research: Learning a first language entails active search for language-specific grammatical patterns (or rules) to express universal cognitive functions. The child may be helped in this by the relative "transparency" (Slobin, 1980) of the speech addressed to it—either because the language itself, like Turkish, is transparent and/or because adult speech to the child is conspicuously well formed. Several studies (e.g., Newport et al., 1977) have shown that the speech addressed to children tends not to be "degenerate." Yet the speech may be "meager" in the sense that relatively few instances suffice to trigger recognition of a pattern (Roeper, 1982). Such rapid learning would seem to require a system specialized for discovering distinctive patterns of sound and syntax in any language to which a child is exposed.

Finally, it is worth remarking that all normal children do learn a language, just as they learn to walk. Western societies acknowledge this in their attitude to children who fail: we regard them as handicapped or defective, and we arrange clinics and therapeutic settings to help them. As Dale (1976) has remarked, we do not do the same for children who cannot learn to play the piano, do long division, or ride a bicycle. Of course, children vary in intelligence, but not until I.Q. drops below about 50 do language difficulties begin to appear (Lenneberg, 1967). Children at a given level of maturation also vary in how much they talk, what they talk about, and how many words they know. Where they vary little, it seems, is in their grasp of the basic principles of the language system—its sound structure and syntax.

CONCLUSION

The past 50 years have seen a vast increase in our knowledge of the biological foundations of language. Rather than attempt even a sampling of the issues raised by the research we have reviewed, let me end by emphasizing a point with which I began: the interplay between basic and applied research, and between research and theory.

The advances have come about partly through technological innovations, permitting, for example, physical analysis of the acoustic structure of speech and precise localization of brain abnormalities; partly through methodological gains in the experimental analysis of behavior; partly through growing social concern with the blind, the deaf, and otherwise language-handicapped. Yet these scattered elements would still be scattered had they not been brought together by a theoretical shift from description to explanation.

Perhaps the most striking aspect of the development is its unpredictability. Fifty years ago no one would have predicted that formal study of syntax would offer a theoretical framework for basic research in language acquisition, now a thriving area of modern experimental psychology, with important implications for treatment of the language-handicapped. No one would have predicted that applied research on reading machines for the blind would contribute to basic research in human phonetic capacity, lending experimental support to the formal linguistic claim of the independence of phonology and syntax. Nor, finally, would anyone have predicted that basic psycholinguistic research in American Sign Language would provide a unique approach to the understanding of brain organization for language and to testing the hypothesis, derived from linguistic theory, that language is a distinct faculty of the human mind.

Presumably, continued research in the areas we have reviewed and in related areas that we have not (such as the acquisition of reading, the motor control and coordination of articulatory action, second language learning), will consolidate our view of language as an autonomous system of nested subsystems (phonology, syntax). Beyond this lies the further task of unfolding the language system, tracing its evolutionary and ontogenetic origins in the nonlinguistic systems that surround it and from which, in the last analysis, it must derive. We would be rash to speculate on the diverse areas of research and theory that will contribute to this development.

* * *

I thank Ignatius Mattingly for comments and advice.

REFERENCES

Aslin, R.N., Pisoni, D.B., and Jusczyk, P.W.
1983 Auditory development and speech perception in infancy. In M.M. Haith and J.J. Campos, eds., *Infancy and the Biology of Development*. Vol. II: *Carmichael's Manual of Child Psychology*. 4th ed. New York: John Wiley and Sons.

Baker, C., and Padden, C.A.
1978 Focusing on the nonmanual components of American Sign Language. Pp. 27–58 in P. Siple, ed., *Understanding Language Through Sign Language Research*. New York: Academic Press.

Bates, E., and MacWhinney, B.
1982 Functionalist approaches to grammar. Pp. 173–218 in E. Wanner and L.R. Gleitman, eds., *Language Acquisition: The State of the Art*. New York: Cambridge University Press.

Bellugi, U., Poizner, H., and Klima, E.S.
1983 Brain organization for language: clues from sign aphasia. *Human Neurobiology* 2:155–170.

Benson, D.F.
1983 Cerebral metabolism. Pp. 205–211 in M. Studdert-Kennedy, ed., *Psychobiology of Language*. Cambridge: MIT Press.

Bever, T.G.
1970 The cognitive basis for linguistic studies. In J.R. Hayes, ed., *Cognition and Language Development*. New York: John Wiley and Sons.

Bloomfield, L.
1933 *Language*. New York: Holt.

Brown, R.
1973 *A First Language: The Early Stages*. Cambridge: Harvard University Press.

Brown, R., and Hanlon, C.
1970 Derivational complexity and order of acquisition in child speech. In J.R. Hayes, ed., *Cognition and the Development of Language*. New York: John Wiley and Sons.

1976 *Psycholinguistics*. New York: Holt,

Caramazza, A., and Zurif, E.B.

1976 Comprehension of complex sentences in children and aphasics: a test of the regression hypothesis. Pp. 145–161 in A. Caramazza and E.B. Zurif, eds., *Language Acquisition and Language Breakdown*. Baltimore: Johns Hopkins University Press.

Carey, S.

1982 Semantic development: the state of the art. In E. Wanner and L. Gleitman, eds., *Language Acquisition: The State of the Art*. New York: Cambridge University Press.

Chiba, T., and Kajiyama, M.

1941 *The Vowel: Its Nature and Structure*. Tokyo: Tokyo-Kaiseikan.

Chomsky, N.

1956 Three models for the description of language. *IRE Transactions on Information Theory* IT-2:113–124.

1957 *Syntactic Structures*. The Hague: Mouton.

1959 Review of *Verbal Behavior* by B.F. Skinner. *Language* 35:26–58.

1965 *Aspects of the Theory of Syntax*. Cambridge: MIT Press.

1972 *Language and Mind*. New York: Harcourt Brace Jovanovich (revised edition).

1975 *Reflections on Language*. New York: Random House.

1980 Rules and representations. *The Behavioral and Brain Sciences* 3:1–62.

Chomsky, N., and Halle, M.

1968 *The Sound Pattern of English*. New York: Harper and Row.

Cole, R.A., and Scott, B.

1974 Toward a theory of speech perception. *Psychological Review* 81:348–374.

Cole, R.A., Rudnicky, A., Reddy, R., and Zue, V.W.

1980 Speech as patterns on paper. In R.A. Cole, ed., *Perception and Production of Fluent Speech*. Hillsdale, N.J.: Lawrence Erlbaum Associates.

Cooper, F.S.

1950 Spectrum analysis. *Journal of the Acoustical Society of America* 22:761–762.

1972 How is language conveyed by speech? Pp. 25–45 in J.F. Kavanagh and I.G. Mattingly, eds., *Language by Ear and by Eye: The Relationships Between Speech and Reading*. Cambridge: MIT Press.

Cooper, F.S., and Borst, J.M.

1952 Some experiments on the perception of synthetic speech sounds. *Journal of the Acoustical Society of America* 24:597–606.

Cooper, F.S., Gaitenby, J., and Nye, P.W.

1984 Evolution of reading machines for the blind: Haskins Laboratories' research as a case history. *Journal of Rehabilitation Research and Development* 21:51–87.

Crain, S., and Shankweiler, D.

In press Reading acquisition and language acquisition. In A. Davison, G. Green, and G. Herman, eds., *Critical Approaches to Readability: Theoretical Bases of Linguistic Complexity*. Hillsdale, N.J.: Lawrence Erlbaum Associates.

Dale, P.S.

1976 *Language Development*. 2nd ed. New York: Holt, Rinehart and Winston.

Darwin, C.J.

1976 The perception of speech. In E.C. Carterette and M.P. Friedman, eds., *Handbook of Perception*. Vol. 7, *Language and Speech*. New York: Academic Press.

Dennis, M.

1983 Syntax in brain-injured children. Pp. 195–202 in M. Studdert-Kennedy, ed., *Psychobiology of Language*. Cambridge: MIT Press.

e Saussure, F.

1966 *Course in General Linguistics* (Translated by Wade Basuin). New York: McGraw-Hill.

e Villiers, J.G., and de Villiers, P.A.

1973 A cross-sectional study of the acquisition of grammatical morphemes. *Journal of Psycholinguistic Research* 2:267–278.

:imas, P.D.

1982 Speech perception: A view of the initial state and perceptual mechanics. Pp. 339–360 in J. Mehler, E.C.T. Walker, and M. Garrett, eds., *Perspectives on Mental Representation*. Hillsdale, N.J.: Lawrence Erlbaum Associates.

'ant, G.

1960 *Acoustic Theory of Speech Production*. The Hague: Mouton.

1962 Descriptive analysis of the acoustic aspects of speech. *Logos* 5:3-17.

1968 Analysis and synthesis of speech processes. Pp. 173–277 in B. Malmberg, ed., *Manual of Phonetics*. Amsterdam: North-Holland.

1973 Descriptive analysis of the acoustic aspects of speech. *Speech Sounds and Features* (Chapter 2). Cambridge: MIT Press.

'lanagan, J.L.

1983 *Speech Analysis, Synthesis and Perception*. Heidelberg: Springer-Verlag.

'odor, J.

1982 *Modularity of Mind*. Cambridge: MIT Press.

'odor, J.A., Bever, T.G., and Garrett, M.F.

1974 *The Psychology of Language*. New York: McGraw-Hill.

'oss, D.J., and Hakes, D.T.

1978 *Psycholinguistics: An Introduction to the Psychology of Language*. Englewood Cliffs, N.J.: Prentice-Hall.

iivon, T.

1979 *On Understanding Grammar*. New York: Academic Press.

ioodenough, C., Zurif, E.B., and Weintraub, S.

1977 Aphasics' attention to grammatical morphemes. *Language and Speech* 20:11–20.

ioodglass, H., and Geschwind, N.

1976 Language disturbance (aphasia). In E.C. Carterette and M.P. Friedman, eds., *Handbook of Perception*. Vol. 7. New York: Academic Press.

ioodglass, H., and Kaplan, E.

1972 *The Assessment of Aphasia and Related Disorders*. Philadelphia: Lea and Febiger.

lecaen, H., and Albert, M.L.

1978 *Human Neuropsychology*. New York: John Wiley and Sons.

lockett, C.F.

1960 The origin of speech. *Scientific American* 203:89–96.

1968 *The State of the Art*. The Hague: Mouton.

akobson, R.

1941 *Kindersprache, Aphasie, und Allgemeine Lautgesetze*. Stockholm: Almqvist and Wiksell.

oos, M.

1948 Acoustic phonetics. *Language Monograph* 23(24):Supplement.

<atz, J.J.

1981 *Language and Other Abstract Objects*. Totowa, N.J.: Rowman and Littlefield.

<imura, D.

1961 Cerebral dominance and the perception of verbal stimuli. *Canadian Journal of Psychology* 15:166–171.

1967 Functional asymmetry of the brain in dichotic listening. *Cortex* 8:163–178.

Kiparsky, P.

1968 Linguistic universals and linguistic change. Pp. 171–202 in E. Bach and R. Harms, eds., *Universals in Linguistic Theory*. New York: Holt, Rinehart and Winston.

Klima, E.S., and Bellugi, U.

1979 *The Signs of Language*. Cambridge: Harvard University Press.

Kurath, H.

1939 *Handbook of the Linguistic Geography of New England* (With the collaboration of Marcus L. Hansen, Julia Bloch, and Bernard Bloch). Providence, R.I.: Brown University Press.

Labov, W.

1972 *Sociolinguistic Pattern*. Philadelphia: University of Pennsylvania Press.

Ladefoged, P.

1980 What are linguistic sounds made of? *Language* 56:485–502.

Lashley, K.S.

1951 The problem of serial order in behavior. Pp. 112–136 in L.A. Jeffress, ed., *Cerebral Mechanisms in Behavior*. New York: John Wiley and Sons.

Lehmann, W.P.

1973 *Historical Linguistics*. New York: Holt, Rinehart and Winston.

Lenneberg, E.H.

1967 *Biological Foundations of Language*. New York: John Wiley and Sons.

Lesser, R.

1978 *Linguistic Investigations of Aphasia*. New York: Elsevier.

Levinson, S.E., and Liberman, M.Y.

1981 Speech recognition by computer. *Scientific American*, April.

Levi-Strauss, C.

1966 *The Savage Mind*. Chicago: University of Chicago Press.

Liberman, A.M.

1957 Some results of research on speech perception. *Journal of the Acoustical Society of America* 29:117–123.

1970 The grammars of speech and language. *Cognitive Psychology* 1:301–323.

1982 On finding that speech is special. *American Psychologist* 37:148–167.

Liberman, A.M., and Studdert-Kennedy, M.

1978 Phonetic perception. Pp. 143–178 in R. Held, H.W. Leibowitz, and H.-L. Teuber, eds., *Handbook of Sensory Physiology. Vol. VIII: Perception*. New York: Springer-Verlag.

Liberman, A.M., Cooper, F.S., Shankweiler, D., and Studdert-Kennedy, M.

1967 Perception of the speech code. *Psychological Review* 74:431–461.

Liberman, A.M., Ingemann, F., Lisker, L., Delattre, P.C., and Cooper, F.S.

1959 Minimal rules for synthesizing speech. *Journal of the Acoustical Society of America* 31:1490–1499.

Licklider, J.C.R., and Miller, G.

1951 The perception of speech. In S.S. Stevens, ed., *Handbook of Experimental Psychology*. New York: John Wiley and Sons.

Liddell, S.K.

1978 Nonmanual signals and relative clauses in American Sign Language. Pp. 59–90 in P. Siple, ed., *Understanding Language Through Sign Language Research*. New York: Academic Press.

Lieberman, P., and Crelin, E.S.

1971 On the speech of Neanderthal man. *Linguistic Inquiry* 2:203–222.

Lieberman, P., Crelin, E.S., and Klatt, D.H.
1972 Phonetic ability and related anatomy of the newborn, adult human, Neanderthal man, and the chimpanzee. *American Anthropologist* 74:287–307.

Limber, J.
1973 The genesis of complex sentences. In T.E. Moore, ed., *Cognitive Development and the Acquisition of Language*. New York: Academic Press.

Linebarger, M.C., Schwartz, M.F., and Saffran, E.M.
1983 Sensitivity to grammatical structure in so-called agrammatic aphasics. *Cognition* 13:361–392.

Locke, J.
1983 *Phonological Acquisition and Change*. New York: Academic Press.

Luria, A.R.
1966 *Higher Cortical Functions in Man*. New York: Basic Books.
1970 *Traumatic Aphasia*. The Hague: Mouton.

MacNamara, J.
1982 *Names for Things*. Cambridge: MIT Press.

Mattingly, I.G.
1968 Experimental methods for speech synthesis by rule. *IEEE Transactions on Audio and Electroacoustics* AU-16:198–202.
1974 Speech synthesis for phonetic and phonological models. Pp. 2451–2487 in T.A. Sebeok, ed., *Current Trends in Linguistics* Vol. 12. The Hague: Mouton.

Mayberry, R.I.
1978 Manual communication. In H. Davis and S.R. Silverman, eds., *Hearing and Deafness* (4th ed.). New York: Holt, Rinehart and Winston.

Mayr, E.
1974 Behavior programs and evolutionary strategies. *American Scientist* 62:650–659.

Miller, G.A.
1951 *Language and Communication*. New York: McGraw-Hill.

Miller, G.A., Galanter, E., and Pribram, K.H.
1960 *Plans and the Structure of Behavior*. New York: Henry Holt and Company, Inc.

Molfese, D.L.
1977 Infant cerebral asymmetry. In S.J. Segalowitz and F.A. Gruber, eds., *Language Development and Neurological Theory*. New York: Academic Press.

Moscovitch, M.
1983 Stages of processing and hemispheric differences in language in the normal subject. Pp. 88–104 in M. Studdert-Kennedy, ed., *Psychobiology of Language*. Cambridge: MIT Press.

Mowrer, O.H.
1960 *Learning Theory and the Symbolic Processes*. New York: John Wiley and Sons.

Müller, J.
1848 *The Physiology of the Senses, Voice and Muscular Motion with the Mental Faculties*. (Translated by W. Baly). New York: Walton and Maberly.

Neville, H.J.
1980 Event-related potentials in neuropsychological studies of language. *Brain and Language* 11:300–318.

Neville, H.J., Kutas, M., and Schmidt, A.
1982 Event-related potential studies of cerebral specialization during reading. II. Studies of congenitally deaf adults. *Brain and Language* 16:316–337.

Newport, E.L., Gleitman, H., and Gleitman, L.R.
1977 Mother, I'd rather do it myself: some effects and non-effects of maternal speech style.

In C. Snow and C. Ferguson, eds., *Talking to Children; Language Input and Acquisition*. Cambridge, England: Cambridge University Press.

Oettinger, A.
1972 The semantic wall. In E.E. David and P.B. Denes, eds., *Human Communication: A Unified View*. New York: McGraw-Hill.

Ojemann, G.A.
1983 Brain organization for language from the perspective of electrical stimulation mapping. *The Behavioral and Brain Sciences* 6:218–219.

Oller, D.K.
1980 The emergence of the sounds of speech in infancy. Pp. 93–112 in G.H. Yeni-Komshian, J.F. Kavanagh, and C.A. Ferguson, eds., *Child Phonology*. Vol. 1: *Production*. New York: Academic Press.

Pick, A.
1913 *Die Agrammatischen Sprachstörungen*. Berlin: Springer.

Porter, R.J., Jr., and Hughes, L.F.
1983 Dichotic listening to CV's: method, interpretation and application. In J. Hellige, ed., *Cerebral Hemispheric Asymmetry: Method, Theory, and Application*. Praeger Science Publishers: University of Southern California Press.

Potter, R.K., Kopp, G.A., and Green, H.C.
1947 *Visible Speech*. New York: D. Van Nostrand Co., Inc.

Pylyshyn, Z.W.
1980 Computation and cognition: issues in the foundations of cognitive science. *The Behavioral and Brain Sciences* 3:111–169.

Reddy, D.R.
1975 *Speech Recognition: Invited Papers Presented at the 1974 IEEE Symposium*. New York: Academic Press.

Roeper, T.
1982 The role of universals in the acquisition of gerunds. In E. Wanner and L.R. Gleitman, eds., *Language Acquisition: The State of the Art*. New York: Cambridge University Press.

Seashore, R.H., and Erickson, L.D.
1940 The measurement of individual differences in general English vocabularies. *Journal of Educational Psychology* 31:14–38.

Segalowitz, S.J., and Chapman, J.S.
1980 Cerebral asymmetry for speech in neonates: a behavioral measure. *Brain and Language* 9:281–288.

Shankweiler, D., and Studdert-Kennedy, M.
1967 Identification of consonants and vowels presented to the left and right ears. *Quarterly Journal of Experimental Psychology* 19:59–63.

Skinner, B.F.
1957 *Verbal Behavior*. New York: Appleton-Century Crofts.

Slobin, D.I.
1980 The repeated path between transparency and opacity in language. In U. Bellugi and M. Studdert-Kennedy, eds., *Signed and Spoken Language: Biological Constraints on Linguistic Form*. Weinheim: Verlag Chemie.

1982 Universal and particular in the acquisition of language. In L. Gleitman and E. Warner, eds., *Language Acquisition: State of the Art*. New York: Cambridge University Press.

Stevens, K.N.
1975 The potential role of property detectors in the perception of consonants. Pp. 303–330 in G. Fant and M.A.A. Tatham, eds., *Auditory Analysis and Perception of Speech*. New York: Academic Press.

Stevens, K.N., and House, A.S.

1955 Development of a quantitative description of vowel articulation. *Journal of the Acoustical Society of America* 27:484–493.

1961 An acoustical theory of vowel production and some of its implications. *Journal of Speech and Hearing Research* 4:303–320.

Stokoe, W.C., Jr.

1960 Sign language structure. *Studies in Linguistics: Occasional Papers 8*. Buffalo: Buffalo University Press.

1974 Classification and description of sign languages. Pp. 345–371 in T.A. Sebeok, ed., *Current Trends in Linguistics*. Vol. 12. The Hague: Mouton.

Stokoe, W.C., Jr., Casterline, D.C., and Croneberg, C.G.

1965 *A Dictionary of American Sign Language*. Washington, D.C.: Gallaudet College Press.

Studdert-Kennedy, M.

1974 The Perception of Speech. In T.A. Sebeok, ed., *Current Trends in Linguistics*. Vol. 12. The Hague: Mouton.

1976 Speech perception. Pp. 243–293 in N.J. Lass, ed., *Contemporary Issues in Experimental Phonetics*. New York: Academic Press.

1983 *Psychobiology of Language*. M. Studdert-Kennedy, ed., Cambridge: MIT Press.

1985 Sources of variability in early speech development. In J.S. Perkell and D.H. Klatt, eds., *Invariance and Variability of Speech Processes*. Hillsdale, N.J.: Lawrence Erlbaum Associates.

Studdert-Kennedy, M., and Lane, H.

1980 Clues from the differences between signed and spoken language. Pp. 29–39 in U. Bellugi and M. Studdert-Kennedy, eds., *Signed and Spoken Language: Biological Constraints on Linguistic Form*. Deerfield Park, Fla.: Verlag Chemie.

Studdert-Kennedy, M., and Shankweiler, D.P.

1970 Hemispheric specialization for speech perception. *Journal of the Acoustical Society of America* 48:579–594.

Templin, M.

1957 *Certain Language Skills of Children*. Minneapolis: University of Minnesota Press.

Von Stockert, T.

1972 Recognition of syntactic structure in aphasic patients. *Cortex* 8:323–335.

Wanner, E., and Gleitman, L.R., eds.

1982 *Language Acquisition: The State of the Art*. New York: Cambridge University Press.

Werker, J.F.

1982 The Development of Cross-Language Speech Perception: The Effect of Age, Experience and Context on Perceptual Organization. Unpublished Ph.D. dissertation. University of British Columbia.

Wilson, E.O.

1975 *Sociobiology*. Cambridge: The Belknap Press.

Yeni-Komshian, G.H., Kavanagh, J.F., and Ferguson, C.A., eds.

1980 *Child Phonology*. Vols. 1 and 2. New York: Academic Press.

Zaidel, E.

1978 Lexical organization in the right hemisphere. Pp. 177–197 in P.A. Buser and A. Rougeul-Buser, eds., *Cerebral Correlates of Conscious Experience*. Amsterdam: Elsevier/North-Holland Biomedical Press.

1980 Clues from hemispheric specialization. In U. Bellugi and M. Studdert-Kennedy, eds., *Signed and Spoken Language: Biological Constraints on Linguistic Form*. Weinheim: Verlag Chemie.

1983 On multiple representations of the lexicon in the brain—the case of two hemispheres.

Pp. 105–125 in M. Studdert-Kennedy, ed., *Psychobiology of Language.* Cambridge: MIT Press.

Zurif, E.B., and Blumstein, S.E.
1978 Language and the brain. In M. Halle, J. Bresnan, and G.A. Miller, eds., *Linguistic Theory and Psychological Reality.* Cambridge: MIT Press.

Visual Perception of Real and Represented Objects and Events

JULIAN HOCHBERG

INTRODUCTION

Experimental psychology started with the study of how we perceive pictures and of the conditions under which one object is an effective surrogate for another (that is, the two objects elicit the same effect). Such study has served the purposes of other disciplines as well, and remains inherently interdisciplinary.

Prior to 1850 the problem was primarily pursued by artists and philosophers, and the conceptual tools were essentially those of physics and geometry. In the classical period, roughly from 1850 to 1950, the primary theoretical concerns were those of neurophysiologists and psychologists. Major applications—in visual prosthesis (e.g., optometry and ophthalmology), the visual media (e.g., photography, print, and eventually television), and the interface between human and machine currently called *human factors*—motivated much of the research that provided a rich base of technical data.

The present period of tremendous ferment started around 1950. The problems of perception continue to engage all the disciplines already mentioned; in addition, computer science is now a major presence in the field, providing tools and motivation in several distinct but closely related ways: as a source of techniques for research, theory testing, and modeling; as a source of analogies and metaphors; as an overlapping enterprise, seeking to devise machines that will "perceive" in the same way that people do; and in the context of learning how to generate and display computer images that humans can readily and accurately comprehend.

THE PRE-1850s: ARTISTS, PHILOSOPHERS, AND PHYSICISTS

Artists have known for centuries that one way to produce a picture is make a surrogate object that (ideally) offers the eye the same pattern light as that offered by the scene itself. The most famous example of t is Leonardo's window (Figure 1A): By tracing the outlines of objects o plane of glass interposed between his eye and the scene, the artist discov the characteristics of a two-dimensional projection of a three-dimensio scene. Of course, the method could be used to provide pictures of existi

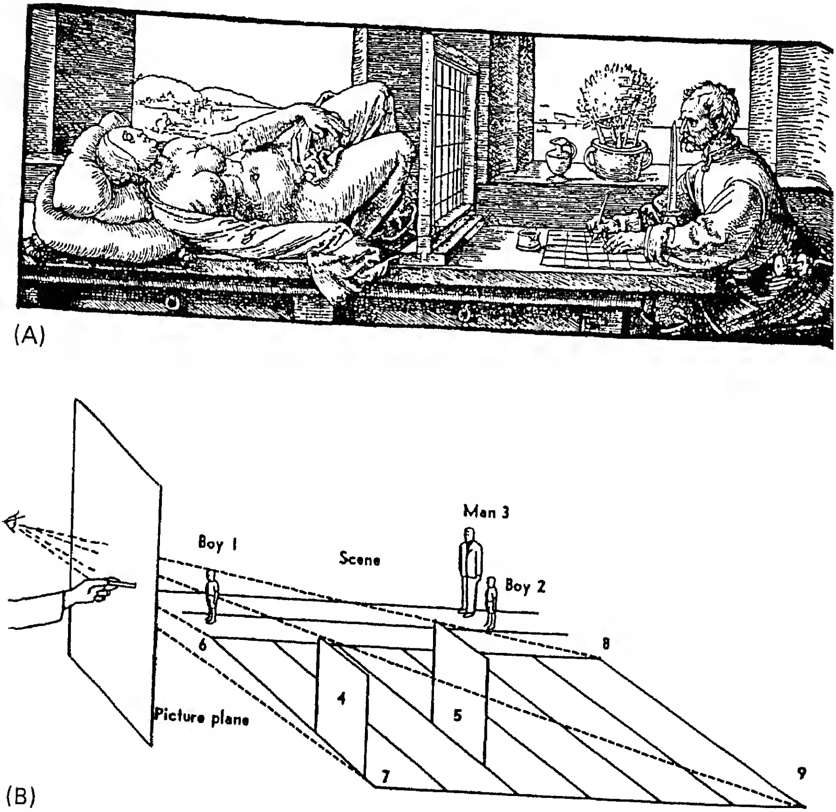


FIGURE 1 Surrogates and their preparation. A: One of the optical aids that artists have used for centuries (Dürer) to help in preparing a surrogate that provides the eye with much of the same stimulus information as the object or scene being represented. B: By studying the tracings made of scenes viewed through a glass pane Leonardo advised that artists could learn the characteristic two-dimensional projections of three-dimensional layouts and could then construct pictures of imagined scenes.

scenes with no need for the artist to learn anything: the scene could be traced directly on the glass (Figure 1B) or—with the growth of technology—by photographic or video media.

Some traditional features that result from projecting normal three-dimensional scenes on two dimensions appear in Figure 2: these include *linear perspective, familiar size, relative size*, and *interposition*. Note that even a perfect picture produced in this way is inherently ambiguous, in that both the flat surrogate and the very different three-dimensional layout it represents offer the same light to the eye. This is the aspect of pictures that made them, and visual perception, of interest to the philosophers—the epistemological issue of how we can know what is true. Philosophical concerns aside, the ambiguity is inherent as a matter of simple mathematics, and provides both the opportunity for pictorial communication and a tool for psychological and physiological inquiry.

The artist who learns to use signs of depth, as in Figure 2, can produce surrogates of scenes that do not and perhaps could not exist—*virtual scenes* of grottos, unicorns, and biblical and extraterrestrial events. Indeed, we shall see that in the interest of visual comprehensibility it is necessary to depart from pure projection, and most pictures are therefore to some extent surrogates of virtual rather than actual scenes.

Today computers provide an increasing proportion of the still and moving pictures that humans confront. For them to do so programmers must learn how to project three-dimensional layouts in two-dimensional arrays and to generate the play of light and shade by which different surface textures are

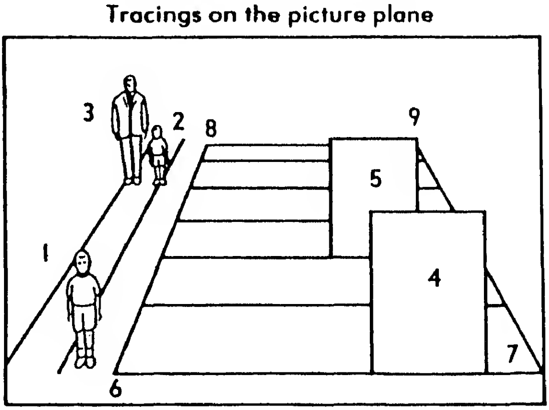


FIGURE 2 The major pictorial (monocular) depth cucs: the tracing of the sccne in Figurc 1B. *Linear Perspective*: parallel lines 6–8, 7–9, etc., convergc in the picture plane. *Interposition*: thc nearer objcct 4 occludes part of thc farther object 5. *Relative Size*: the tracing of boy 1 is larger than that of boy 2. *Texture-Density Gradient*: the evenly spaced bars on the field 6–7–8–9 project an image whose density incrcases with distance. *Familiar Size*: if man 3 is known to bc larger than boy 1, and they are thc samc sizc in the picture plane, then the man must be proportionally farther away in the reprcsented scene.

Surrogates are therefore more than means of pictorial communicatio they tell us about the limits of the information that the sense organ can pick up and about how the brain organizes that information. Perhaps the earlies major instance of that point was in Newton's (1672) famous experiment i visual sensation, showing that an appropriate mix of three narrow wavelengths of light—bands of color taken out of the spectrum, such as red green, and blue—can serve as a surrogate for any and all colors in the spectrum, and thus match any scene (Figure 4). *This is not a fact abou photic energy*—the light itself remains unchanged by the mixture. It i instead a strong clue about our sensory nervous systems, and it provide the background for the classical theory of perception and the nervous system which we consider next.

PSYCHOLOGY AND PHYSIOLOGY FROM 1850–1950

Given the facts of color mixture, the most parsimonious model of visua perception was the Young-Helmholtz theory (Helmholtz, 1866): that colo perception is mediated by three kinds of specialized receptor neurons, th *cones*, each responsive to most of the spectrum, but each with a differer sensitivity function. The three types were thought to be most sensitive t light that looks red, green, and blue, respectively, and their response t

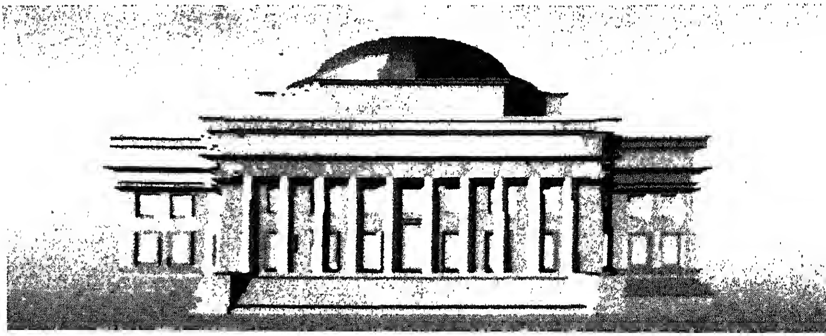


FIGURE 3 Computer-drawn image. A picture programmed directly from blueprin of a building, using a polygon facet approach with a simple lighting model that simulat direct sun and diffuse sky illumination. Paul Roberts, Computer Vision Lab, Columb University.

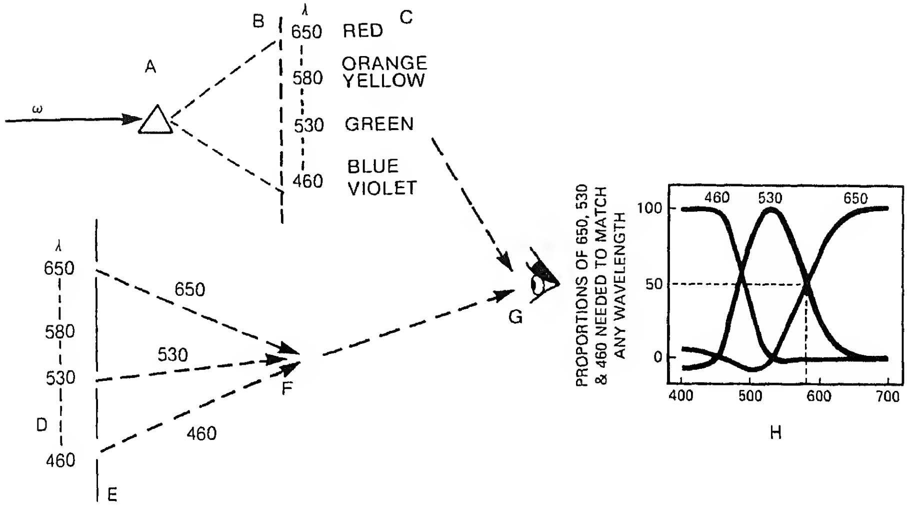


FIGURE 4 Color surrogates. To the visual system G, a suitable mixture F of just three wavelengths selected by slits E from the visible spectrum D can, in principle, be a surrogate for any hue C, that is, any set of wavelengths in the spectrum. According to the traditional Young-Helmholtz theory, the physiological explanation involves three types of retinal cone cells with the three sensitivity functions shown in H. From these we can see, for example, that a mix of equally effective intensities at 650 and 530 is indistinguishable from 580 and could serve as a surrogate for the latter.

photic energy was thought to underlie the experience of those colors. The retina was envisioned as a mosaic of independent triads of the three cones (Figure 5), and the light provided to the eye by any scene was thought to be analyzed into the point-responses of the three component colors. The research most directly relevant to this theory was the attempt to map the sensitivity of each type of cone to the wavelengths of the visible spectrum and to map the spatial resolution of the retinal mosaic—what detail the eye could be expected to resolve.

Such information as the limits of resolution and the bases and specification of colors provided the first goals for what has become *visual science* and its applications, which now run from the prescription of spectacles to the design of television characteristics. It was also the foundation of the classical view of the perceptual process in general, diagrammed in Figure 6: at left, the object in the world, with its physical properties of distance, size, shape, reflectance (surface color). These do not affect the sense organs directly, of course, but only by means of the light they reflect to the sensitive cells. All things that cause the cells to respond in some specific way elicit the same sensory experience: the light coming from the object itself, the light produced by some surrogate of that object, the effects of mechanical or

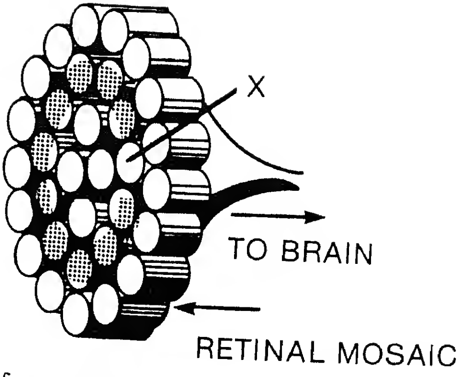


FIGURE 5 The sensory mosaic. In the simplest view, the retina of the eye contains a mosaic of light-sensitive cells. The spacing of the mosaic determines what detail can be seen: e.g., to distinguish a "C" from an "O," at least one cell (x) must go unstimulated. The visible portion of the electromagnetic radiation incident at each point in the retina that is capable of full color vision is coded into the output of each of three cones according to its sensitivity curve (Figure 4H). This is, of course, essentially the way in which video cameras analyze the light they receive from scenes.

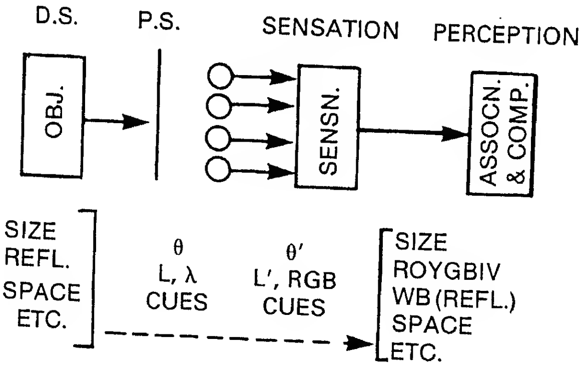


FIGURE 6 The classical theory (1850–1950). The *distal stimulus*, D.S. (an object or layout of objects), with such physical properties as size, reflectance, position in space, etc., impinges on the sensory surface by way of the *proximal stimulus* pattern, P.S., consisting of regions that vary in their spatial extent (θ) and spectral distribution [luminance (L), wavelengths (λ)]. Sensory responses to each region (sensations) were thought to vary correspondingly in brightness (L′) and hue (the mix of Red, Green, and Blue) over some extent (θ). Because of the regularities of the world and its geometry, the proximal stimulation will generally contain patterns (e.g., the cues in Figure 2) that are characteristic of and therefore provide information about the distal properties. The perception of such properties (objects' sizes, surface reflectances, spatial location, etc.) were thought to derive from the underlying sensations by associative learning and by computational processes.

electrical stimulation of the eye, etc. Insofar as different objects and events produce the same responses, information about the world is lost in this *encoding* process. This is what makes surrogates possible. And the fact that very different objects, and indeed different patterns of light, have the same effects on the nervous system provides a tool with which to study that system's structure and function.

The visual system thus conceived is a mosaic of receptors (the retina) on which the eye's optical system projects a focused image of the light provided by the object. The receptors (the three types of cone, supplemented by rods, which do not differentiate color) analyze each small region of that image into points of red, green, and blue. This conception of the visual system has now been embodied in the television camera: Television, like the Helmholtzian visual system, reduces the countless objects and events of the world to the different combinations of a set of three colors in a spatial mosaic. It is important both for the Helmholtzian theory and for television as a medium that such a simple set will suffice. In both cases, all of the remaining properties of the objects that we perceive in the world—their sizes, forms, and reflectances (i.e., surface colors), their distances and movements—are lost in the encoding process and must be supplied by the viewer.

The simplest theory about such nonsensory processes was inherited from centuries of philosophical analyses of perception: the theory that we have learned the perceptual properties of objects from our experiences with the world. It runs as follows:

The sense organs analyze the world into *fundamental sensations*.

Those sensations are, in the case of vision, the sets of points that differ in hue (R, G, B in Figure 6, signifying red, green, and blue sensory experiences) and brightness (L′) over some effective extent, θ′. These packets of sensations normally come in characteristic patterns that are imposed by the regularities of the physical world, patterns such as the depth cues in Figure 2. By learning these regularities and their meanings, we learn to perceive the physical world and its properties.

The theory seems to be economical and elegant. The principles of learning appeared to be at hand. For almost two centuries (from Hobbes in 1651 to James Mill in 1829), the British empiricist philosophers had discussed how the "laws of association," offered in essence by Aristotle, could serve to build our perceptions and ideas about the objects and events of the world. And a plausible neurophysiological explanation of association readily offered itself in terms of increased readiness of nerve cells that had been repeatedly stimulated simultaneously to fire together.

This outline of how we perceive objects and their pictures fitted nicely into a general theory of knowledge and of science, spanning from neuro-

physiology to sociology and political science. With respect to the last,
example, the view that all our ideas about the world derive from
experiences with it leads readily (but not ineluctably) to the belief
human intelligence and character are generally perfectible through edu
tion, and to the advocacy of egalitarianism and individualism over a w
range of social and political issues.

Although formulated by Helmholtz only one academic generation a
his teacher, Johannes Mueller (1838), first undertook the scientific anal
of experience, what I am calling the classical theory of perception thus
wide and deep connections with the mainstream of Western thought,
it remained the dominant theory in neurophysiology and psychology u
the 1950s.

It was not without opposition, however. Some opposition was base
a cluster of purely psychological flaws. Although, for example, the th
tells us which different stimuli will act as mutual surrogates—that is, w
different objects will produce the same perceptual experience—it does
tell us what that experience will be like. It does not predict the attrib
of the experience itself, i.e., it tells us that light composed of a mixtur
650 nanometers (red) and 540 nanometers (green) is indistinguishabl
appearance from light of 580 nanometers (nm) (yellow), but it gives u
basis for predicting how that appearance is similar to and different f
other colors. As we will see, alternative theories, almost as old as
Helmholtzian one, offer much more in the way of accounting for app
ance. Notable among these proposals based on *phenomenology* (the st
of appearances as such) were the following: Hering (1878) argued
perceived colors comprise red-green, yellow-blue, and black-white o
nent systems; that connections between cells of the two retinas provid
an innate sense of depth; and that lateral inhibition between adjacent reg
of the visual system make their appearances mutually dependent. M
(1886) proposed (among other things) that such lateral connections pro
networks that are sensitive to contours and not merely to incident ene

A related problem is illustrated in Figure 7. In most situations in the
world, the local stimulation that is projected to the eye is not by i
information about object properties. Even if the two gray target disk
the cube are of identical lightness or *reflectance* (R_t), the *luminanc*
photic energy each provides the eye is different (L_1, L_2) because the
mination falling on each is different (E_1, E_2). Again, even if the two ver
rods on the right are of the same physical size (S), the size of the re
image each provides (θ_1, θ_2) differs because the rods lie at different dista
(D_1, D_2). Nevertheless, we tend to perceive such object properties corre
despite changing retinal stimulation. The classical theory held that
object constancy, as it is now known, is achieved when the viewer t

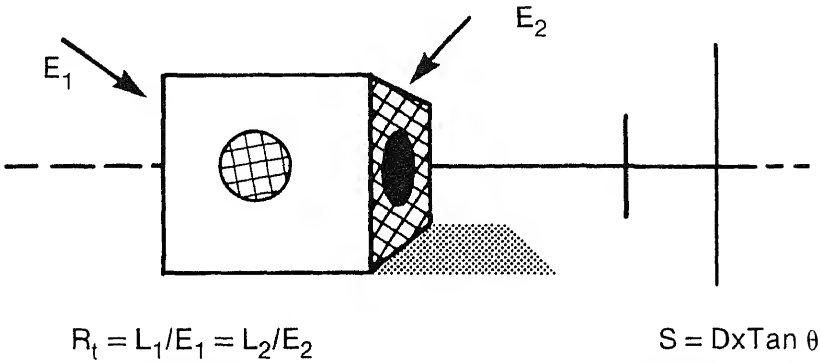


FIGURE 7 Object constancy. Although both target disks on the cube have the same reflectance (R_t), the luminances (L_1, L_2) differ to the eye because the illuminations (E_1, E_2) differ. Similarly, objects of the same size (S) provide images of different extent (θ_1, θ_2) depending on their distances (D_1, D_2). We tend to see objects' relatively permanent qualities, such as their reflectance and size, as constant even though the proximal stimulation they provide is in flux. In the classical theory we do this by learning to process visual information according to the formulae R(reflectance) = L(luminance)/E(illumination), and S(size) = D(distance) × tangent of θ(visual extent).

the conditions of seeing into account: in effect, by using the depth cues to perceive depths D_1 and D_2, and then using the latter to infer the object sizes from the retinal sizes (θ_1, θ_2); similarly, to use cues to perceive the illuminations E_1 and E_2, and, using the latter, to infer the reflectances of the parts of the scene from their luminances.

This explanation is now commonly called "unconscious inference." Its operation assumes that the viewer has learned the constraints in the physical world (e.g., that $L = R \times E$, that $S = kD\tan\theta$, etc.). These constraints, once learned, provide a mental structure that mirrors the physical relationship between the attributes of the object and those of sensory stimulation, permitting the viewer to infer or compute the former from the latter. A general form of this explanation is that *we perceive just that state of affairs in the world that would, under normal conditions, be most likely to produce the pattern of sensory responses we receive.*

The learning processes that might underlie such computations have never been formally and explicitly worked out. What we would now call "lookup tables" (for example, with grouped entries for S, θ, and D) would be compatible with theories about associative learning. Helmholtz and others often wrote, however, as though we learn to apply the *rules* that mirror those of the physical world; they did not say explicitly, however, how such abstract principles, as distinguished from lookup tables listing the elements of sense data, are learned.

The Helmholtzian idea that our perceptions of objects rest on computational or inferential process was, like the classical theory's failure to predict appearances, roundly criticized over the years as being uneconomical, mentalistic, and unparsimonious. Gestalt theory, which had a significant impact in psychology and art theory between the two world wars, was particularly vocal in this regard. But the criticisms of the classical theory did not amount to much until the end of World War II. Then the needs of new technology (flight training, radar and sonar displays, etc.), the development of new instrumentation (notably, direct amplifiers that made the measurement of very small bioelectrical tissue responses common and reliable), and the effects of grants that made the research career a viable occupation, all combined to turn the tables. As we will see, Helmholtz was right about the three cones and in some sense about the existence of mental structure and computation. But most of the rest of what lay between those points was wrong, and most of the alternative proposals that had been made by the critics of that dominant approach, especially those of Hering and Mach, were quite remarkably vindicated within a period of a very few years, after having been largely ignored for many decades.

THE 1950s AND AFTER: "DIRECT" SENSITIVITY TO OBJECT ATTRIBUTES

The two main arguments on which the classical theory rested were, first, that it was the simplest answer to the problem of analyzing the world of sensory stimulation, and second, that it was in accord with neurophysiological observation. In the 1950s both of these supports were withdrawn.

Technically, as is widely recognized, the most important single advance in instrumentation was the microelectrode, which made it possible to record the activity of individual nerve cells in the visual system and brain of an essentially intact animal that is exposed to various sensory displays. It quickly became evident that most of the cells observed in this way respond not to individual points of local stimulus energy but to extended spatial and temporal patterns—to adjacent differences in intensity, specific features, and movements in one rather than another direction (Figure 8). They appear to do so by means of networks of lateral connections, which were very much what Hering and Mach had argued.

In the 1950s Hurvich and Jameson (1957) offered sensitivity curves for the red-green and yellow-blue opponent process cells that Hering had proposed, using procedures based on colors' appearances and not just on their discriminabilities (Figure 9A).

They "titrated" the response that each of these hypothetical red-green and blue-yellow opponent pairs makes to wavelengths throughout the spec-

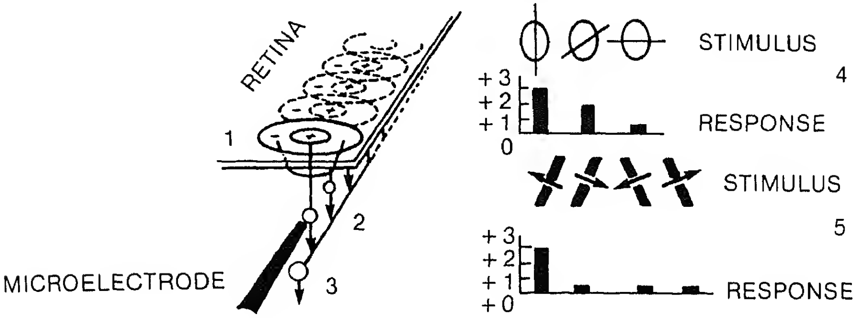


FIGURE 8 Pattern-sensitive neural network. Microelectrode recordings from individual cells in the visual system (Hartline, 1949; Hubel and Wiesel, 1962) reveal far more complex organization than the simple individual punctate analysis of Figures 5 and 6. For example, receptors in the retina 1 send both excitatory (+) and inhibitory (−) connections to more proximal cells 2; those connections are arranged in networks so that cells at level 2 are stimulated by light falling in a center region and inhibited by light falling in its surround. Other cells 3 still deeper in the system are so connected as to be more highly stimulated by a bar or edge falling on the line of 2 than by bars of other orientation 4. Cells farther in the nervous system are sensitive to a bar of specific orientation moving in a specific direction 5.

trum by determining how much of each pure hue was needed to cancel all traces of its opponent. That is, how much pure red was needed to cancel the greenishness at each point between approximately 480 and 580 nm, thus indicating the strength of the response labeled G; how much pure green would cancel the reddishness of wavelengths above and below this region, thus indicating the strength of the response labeled R; etc. Wavelengths that appear as blue, green, yellow, orange, and red are shown at 1–5, respectively, and what they look like can be read off the graph.

In explanation of these functions, Hurvich and Jameson (1974) proposed the following networks (see Figure 9B): Given three cones with the sensitivity functions they are now known to have (only approximately those of Figure 4H) and the network of excitatory connections (solid lines) and inhibitory connections (dotted lines) that is shown, each of the rightmost cells would serve as one or another of the yoked opponents by firing above or below their baseline activity.

Informed by the opponent-process theory, microelectrode research identified cells in the visual system of the goldfish (Svaetichin, 1956) and in the rhesus macaque (DeValois and Jacobs, 1968) that responded to wavelength in just these ways.

Moreover, cells have been found that respond to lines and edges, at particular orientations, moving and stationary (Hubel and Wiesel, 1962),

A B

FIGURE 9 Accounting for color appearance by opponent-process networks. The functions of the Young-Helmholtz theory in Figure 4H explain why three wavelengths suffice to match any color, but do not explain color appearance. Hering had proposed two kinds of units, one that responds with a red sensation to some parts of the spectrum and with a green sensation to others, and a second that responds either blue *or* yellow. These units, by their combined activity, would account for the appearances of all hues. Hurvich and Jameson (1957) charted the amount of these components in the appearance of each section of the spectrum (see text), suggesting the functions in (A) as the response curves of the two kinds of unit, and suggesting a simple network (B) to encode the responses of the three kinds of cone (α, β, γ) into the opponent process hues plus black and white (Hurvich and Jameson, 1974). Anticipated and guided by these analyses of perceptual experience, opponent process cells have been identified and studied by neurophysiological means (Svaetichin, 1956; DeValois, 1968).

to what may be thought of as sine-wave gratings of a particular frequency (Blakemore and Campbell, 1969), to disparities in the two eyes' views (Barlow et al., 1967), etc. Even though the Helmholtzian model (Figures 4–6) may be the simplest, we must conclude that it does not accord with the neurophysiological facts.

These new neurosphysiological structures raise two questions: How do they themselves work, and what perceptual functions do they serve?

With respect to how these structures work, they are widely believed to result from the activities of suitably interconnected networks of lateral inhibition and excitation (von Bekesy, 1960; Ratliff, 1965), like the sketches in Figures 8 and 9; this was very much what Mach and Hering had speculated to be the case.

With respect to their possible perceptual functions, such pattern-sensitive networks open the way to very different kinds of explanation of the perceptual process. One of these is that the visual stimulus is analyzed into fundamental elements that do a great deal of what had been considered the task of learning and of unconscious inference. Three examples that have been given a great deal of attention will be mentioned and must stand for a larger number of such proposals. The first is that our visual world might

be assembled out of such fragments as edges and corners, providing a sort of feature list of which all scenes must be composed. The physiological mechanisms for such analyses could be provided by receptive fields in the striate cortex of the brain that are responsive to lines of a particular orientation anywhere within a local region of the retina (Hubel and Wiesel, 1962, 1968), with cells in the inferotemporal cortex responding to primitive shapes—or even faces—as stimuli, independent of position or orientation (Gross and Mishkin, 1977; Perrett et al., 1982).

A second class of alternatives is to find neural structures that respond directly to specific properties in sensory stimulation that are themselves directly correlated with the distal, physical properties of objects in the world. Thus, cells that are sensitive to a disparity in the two eyes' images might provide a visual mechanism (Barlow et al., 1967) that is directly sensitive to an object's distance, as Hering originally argued. This possibility can be entertained, however, only to the very limited degree (see Gogel, 1984) that binocular space can be considered in such point-by-point fashion; in general, we must deal with extended patterns of stimulation and therefore with spatially organized and extended neural mechanisms.

Spatially organized and extended neural structures are exemplified in a third class of alternatives that is based on the following idea of *spatial-frequency channels*: A sine-wave grating is a set of dark and light bars in which the intensity of the light varies in a sine wave. The width of the bars in such a grating defines its spatial frequency (i.e., the number of bars or cycles per degree of visual angle): high spatial frequencies mean fine detail, and the light-to-dark ratio (or contrast sensitivity) needed to discern the bars of each frequency characterizes the acuity of the visual system in terms that are compatible with those used to evaluate television transmissions and displays (Schade, 1956).

But such spatial frequencies are more than just a useful engineering measure. Because the rings of lateral inhibition that surround each stimulated point in the peripheral visual system come in different sizes, cells in the visual system are differentially responsive to spatial frequency. Cells have been found in the cortex that respond electrophysiologically to a particular range of frequencies within a restricted range of orientations in the retinal image (Movshon et al., 1978; DeValois et al., 1976). Moreover, in rough correspondence to these facts, viewers' abilities to detect combinations of sinusoidal gratings (Campbell and Robson, 1964; Graham and Nachmias, 1971), and the aftereffects of exposure to a particular grating (Blakemore and Campbell, 1969; Pantle and Sekuler, 1968), both suggest that different spatial frequencies are being processed by separate *channels*. The relationship between such channels and the physiological finding of specialized response is not clear, nor is it clear what perceptual function,

tion, but their actual role in the perception of objects and events remains in question (see recent reviews by Braddick et al., 1978; Cavanagh, 1984; Foster, 1984; and Graham, 1981).

Many of the present studies searching for the mechanisms of sensory analysis depend on the use of microelectrodes, but units of sensory analysis much like these had been investigated long before the microelectrode was developed. For example, by showing that prolonged exposure to a particular stimulus event provides the kind of aftereffect that one would expect to find if a receptor were depleted or "fatigued" by that exposure, an argument could be made for the sensory nature of the response to the event. Thus, after exposure of the receptor to a set of horizontal stripes moving continuously downward, a stationary set of such stripes appears to move upward, supporting the argument that the perception of movement rests on a direct sensory response to motion (Wohlegemuth, 1911). This method has proliferated in recent years (see Graham, 1981; Harris, 1980), but such findings can be interpreted in other ways, and the search for new sensory units received greater legitimacy from the neurophysiological findings.

If we change what we take to be the units of sensory analysis, then what we attribute to more central processes must in general change as well. Of greatest theoretical significance are those sensory mechanisms whose output remains invariant even though the local stimulation at each point on the retina may vary, i.e., mechanisms that respond to aspects of the stimulation that covary directly with the physical properties of objects and events. For example, the frog's retina contains cells that respond not to the intensity of light in some part of the retinal image, but to the *ratio* of intensities of surrounded and surrounding regions (Campbell et al., 1978). As has been realized since Hering and even Helmholtz, that ratio remains invariant regardless of changes in illumination—as long as both regions are equally illuminated—so that as sketched in Figure 10 equal ratios of luminance in the proximal stimulation (P.S.) signify equal ratios of reflectances between the object and its background as distal stimuli (D.S.). It has been argued therefore that our perceptions of lightness are responses to adjacent ratios of luminance (Wallach, 1948). Such mechanisms might explain the constancies directly, that is, no additional process of computation or inference need be postulated. They therefore make possible very different explana-

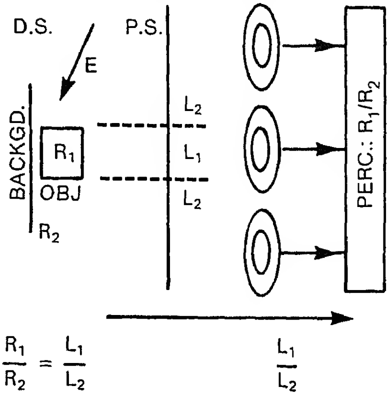


FIGURE 10 A direct response to an object's reflectances. Given that visual neurons are organized into networks, alternative explanations of perception, very different from the classical theory, become possible. For example, if there are networks directly responsive to adjacent ratios of luminance (L_1/L_2), then direct response to an object and its background whose reflectances stand in the ratio R_1/R_2 would remain constant regardless of changes in illumination, E.

tions of how a given visual attribute of objects (color, size, form, distance, velocity) is perceived, explanations that need not draw on speculations either about learning or about computation.

Indeed, because such proposals are useful as perceptual theories only insofar as they identify some aspect of stimulation that "specifies" (i.e., that is highly correlated with) some object property, they need not even be concerned with neurophysiology. The search for such *directly informative variables of stimulation* therefore actually antedates the neurophysiological discoveries (Gibson, 1950) and remains an influential approach today.

The most sweeping and radical proposal of this kind is a direct theory for all of perception (Gibson, 1966, 1979): Our nervous systems "resonate" to stimulus properties that remain invariant when the light at the eye undergoes transformations (e.g., the optical flow patterns and motion parallax, Figure 11) due to relative motion between the viewer and the objects being viewed.

This is of course very different from the traditional approach. The latter posed the original perceptual problem as this: How are we to account for the objects and layouts we do in fact perceive, given that the light at the eye is ambiguous and can be provided by very different surrogates? And it solved that problem by appealing to associations and computations that the individual perceiver has learned from experiences with the world. To the earlier direct theories that opposed this answer (including those of Hering and Mach) and that aimed at explaining particular perceptual abilities, evolution has provided specific mechanisms that so constrain the viewer's responses that they will usually be the correct solution. Some of the newer direct theories seek a much more general principle and are therefore not to be identified with some specific physiological mechanism.

The "invariance" principle is the most general explanation of this kind

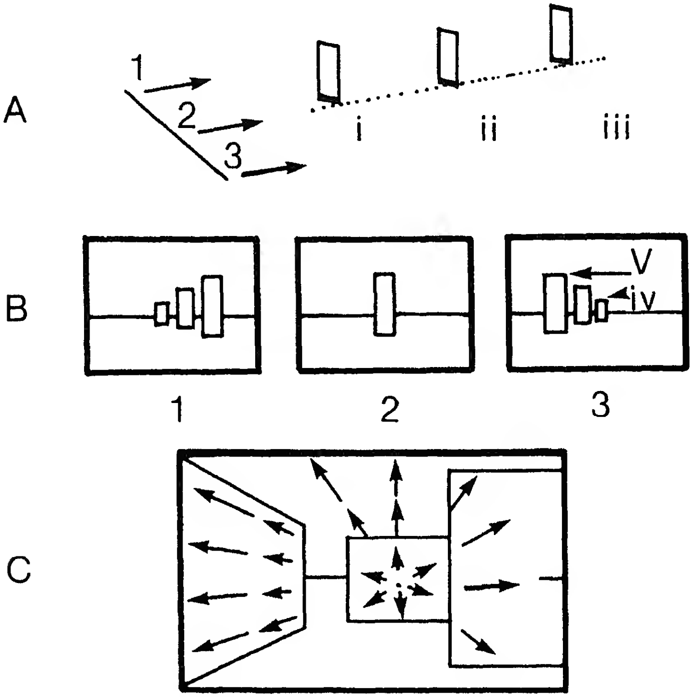


FIGURE 11 Information about layout provided by motion. The views that an observer moving from point 1 to point 3 in A would have of three fixed posts i, ii, iii are shown in B. The motion parallax in those views provides information about the objects' spatial layout and sizes. For example, although the same objects at different distances provide images of unequal size, and are displaced by different amounts (vectors iv, v in B3), the ratio of image size to parallactic displacement should be *invariant*. Gibson (1951, 1966) has emphasized several ways in which the changing pattern of light to the moving observer, such as the *optical expansion patterns* in C, provide potentially usable information about spatial layout and offer invariants that, if responded tò directly, might explain the perception of distal object properties.

to have been offered: Most objects and parts of the environment do not themselves change in form (as smoke or fog do), i.e., are rigid. When applied to these cases, the invariance principle means that we perceive those unchanging, rigid shapes and layouts in the world that project the changing, nonrigid two-dimensional patterns of light to the eye. This assumes that our nervous systems perform the required "reverse projective geometry" (Johansson, 1980), and that wherever the projected light at the eye permits a rigid source to be perceived, it will be.

Because such theories can only account for perception obtained by moving observers, they take the perception of still objects and pictures to be a special case, governed by special and unknown principles (Gibson, 1951, 1979; Johansson, 1980). In this view, normal perception occurs only when an observer moves about in a natural environment; research done in other situations is artificial and therefore misleading about the nature of our perceptual systems as they have evolved.

Both within and outside of this approach, this *rigidity principle* has recently become quite popular. Directly or indirectly (in the form of the assertion that we perceive the invariant), objectwide or more locally, a rigidity principle has been adopted by many psychologists (Gibson, 1966, 1979; Johansson, 1977, 1980; Rock, 1983; Shepard, 1981; Todd, 1982) and computer scientists (Marr, 1982; Ullman, 1979). There are at least three reasons why this principle is theoretically attractive. Exploring those reasons, and why the rule must nevertheless be rejected in any strong form, will provide a convenient survey of a critical part of the present landscape of perceptual inquiry.

The Evidence for Perceptual Rules Rather Than Lookup Tables

It is easy to see how learning by association might invest specific patterns of stimulation with specific perceptual meanings, and to speculate about a neurophysiological basis for such associative learning, but it is harder to be specific about a learning process through which abstract rules might be learned. (This is the distinction, made earlier, between "lookup tables" and an inference or computational process that solves some internalized formula). Criticisms of the classical theory are often simply demonstrations intended to show that perception is determined by rules rather than by familiar associations, rules that operate without, or even against, familiar patterns.

This was the central thrust of Gestalt theory, which mounted a serious challenge to Helmholtzian theory between the two world wars—to find such rules, and from them to deduce the nature of the underlying brain processes. These rules, called the "laws of organization," were held to determine whether we will perceive some object at all (Koffka, 1935; Kohler, 1929). Figure 12A is a demonstration of the "law of good continuation": a familiar number is concealed in i—but not in ii and iii—because the configuration in i requires us to break the unfamiliar but smoothly continuing shape in order to see the number. These rules were also held to determine whether flatness or tridimensionality is perceived (Kopfermann, 1935). In Figure 12C, the pattern looks flat because the good continuation must be broken to perceive (1) and (2) as dihedrals at different distances,

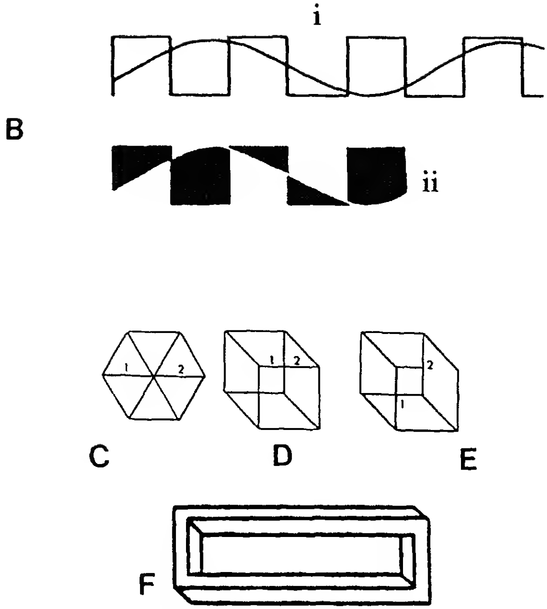


FIGURE 12 Organization and its limits. Before microelectrodes showed extensive cross-connections to exist, similar interaction had been postulated by Gestalt theorists to explain ''laws of organization'' as demonstrated in (A) and (B).

(A) *Good continuation*: a number is concealed in (i) by the smoothly continuing lines that embed it (ii), but not by mere clutter (iii). (B) *Gestalt factors in conflict.* In (i), we perceive a sine wave crossing a square wave, against the factor of *closedness*, which would otherwise yield the perception of closed shapes (ii). (C,D) By the *minimum* principle—that we see the simplest organization—(C) looks flat and (D) looks tridimensional because (C) is simpler as a flat pattern than (D).

(E,F). The evidence is against such global organization. While you gaze at intersection (1) in (E), the vertical line soon appears nearer than the horizontal, which is inconsistent with the simple figure fixed by intersection (2), and does so even with a moving, tridimensional cube (Peterson and Hochberg, 1983). (F) An impossible, yet apparently tridimensional, picture.

but Figure 12D looks tridimensional because the dihedrals would have to be broken at (1), (2), etc., for the pattern to look like a set of closed polyhedra.

On a practical level, such demonstrations remind us that we cannot assume that a picture will be comprehensible if only it is an accurate projective surrogate (as in Figure 1) and as long as the object represented is itself a recognizable and familiar one. Every amateur photographer learns that flowerpots and lampposts lurk in the background, ready to appear in the picture looking very much as though they are growing out of the sitter's head. And any text on protective coloration shows the striped tiger or zebra disappearing into its cluttered background.

On a theoretical level, such demonstrations have been used to argue that associative learning does not determine perception: In Figure 12Ai specific familiarity is overcome by what seems to be an abstract configurational rule.

The literature contains a large number of Gestalt rules, but each is supported only by a few unquantified and untested demonstrations. Nor have the rules been used to explore brain processes. But they do appear to be of the utmost importance inasmuch as they seem to determine what shape or object will be perceived. Because several Gestalt rules usually apply in any real case, however, and because they will as likely as not work against each other, they are not of much use in their present state, lacking quantitative measurement and with no combinatorial rules of any kind. It is not true, as some computer scientists and neurophysiologists have claimed, that these rules have been abandoned because they were inherently subjective and unverifiable (e.g., Marr, 1982). They stand neglected rather than abandoned. The fact is that, until recently, only a handful of scientists were concerned with the problem of organization, and they were deflected by two more promising lines of attack on that problem which seemed to offer themselves in the 1950s.

The Promise of a Minimum Principle To make the insights of Gestalt psychology scientifically or practically useful we need either a great deal of quantitative and object measurement of the strengths of the different rules, along with an appropriate combinatorial principle, or some equally quantitative and objective overarching rule that supplants the set of individual rules. For the latter purpose Gestalt psychologists offered a *minimum principle*, i.e., that we perceive the simplest organization—the simplest alternative object or arrangement—that fits the stimulus pattern (Koffka, 1935). Attempts were initiated in the 1950s (Attneave, 1954, 1959; Hochberg and MacAlister, 1953) to formulate an objective minimum principle, one that would require no intuitive judgments in order to apply it. It would rest instead on measuring each of the alternative objects that could fit the stimulus, to decide which alternative is simpler (e.g., number of dihedrals or edges, number of inflection points, etc. [Hochberg and Brooks, 1960]). With an objective and quantitative rule of this sort, a computer could *in principle* assess any picture before displaying it, and then select for display only those views for which the object to be represented is in fact the simplest alternative (e.g., Figure 12D rather than 12C).

Although no computer programs that would apply these principles to image generation have actually been attempted to my knowledge, development of this approach continues today (Buffart et al., 1981; Butler, 1982; Leeuwenberg, 1971), and it has recently been applied as well to the perception of simple ambiguous patterns of moving dots (Restle, 1979). Such research would be theoretically important if a minimum principle were in prospect and practically important even if all it did was contribute to solving the problems of object representation. But both its theoretical and practical meaning must be questioned in view of facts that have been known to perceptual psychologists for decades. These facts tell us that stimulus measures alone cannot provide a general explanation or prediction of object perception. This will receive increasing stress in the balance of this paper. Here we note that in Figure 12E (p. 266), the place that one attends determines how the object is perceived: when one attends intersection (2), the cube is so perceived that the vertical edge is the nearer, in accordance with both the rule of good continuation and with any simplicity principle; when one attends intersection (1), the perspective soon reverses, against the good continuation at the other intersection and against overall simplicity (Hochberg, 1981).

Both real and pictured objects exhibit this phenomenon (Gillam, 1972; Peterson and Hochberg, 1983). These demonstrations introduce us to the fact that the viewer's attention, and not merely the measurable pattern of stimulation, helps determine what is perceived. (We will return to this point shortly.) With respect to the minimum principle, Figure 12E is completely incompatible with any rule based on the *entire* object. On the other hand, it is not evident how a minimum principle based on *separate parts* of an object can even be formulated and tested. In any case, no advocate of the application of the minimum principle to entire figures has yet attempted to deal with this problem, despite the fact that it was clearly implied by discovery of the famous "impossible figures" by Penrose and Penrose in 1958 and by their popularization in the graphic art of Maurice Escher. The object in Figure 12F (Hochberg, 1968) appears tridimensional and continuous, even though careful inspection of the two sides shows them to be inconsistent. If the distance between left and right sides is made very short, the figure then becomes flat, and the inconsistency more evident, although the minimum principle is then no more or less applicable.

Let us next consider the other factor that deflected attention from the objective study of organizational principles, the assumption that they applied only to stationary drawings.

The Doctrine That Event Perception Is Both Fundamental and Veridical As Leonardo noted in the fifteenth century, a two-dimensional picture cannot provide a moving viewer with the motion parallax that would be provided by the three-dimensional scene it represents. As the viewer moves, nearer objects in a three-dimensional scene are displaced more in the field of view than are farther ones (Figure 13). Because the spatial relationships between the parts of the flat picture all remain fixed, the picture is no longer a surrogate for the scene. The relative motions produced by a given displacement are (with certain constraints or assumptions) specific to the layout of the points and surfaces of the scene in space. The differential motions within the stimulus pattern offered by the scene provide the moving observer with rich information about the structure of the world. A critical question being explored today is how much of that information is used, and in what form.

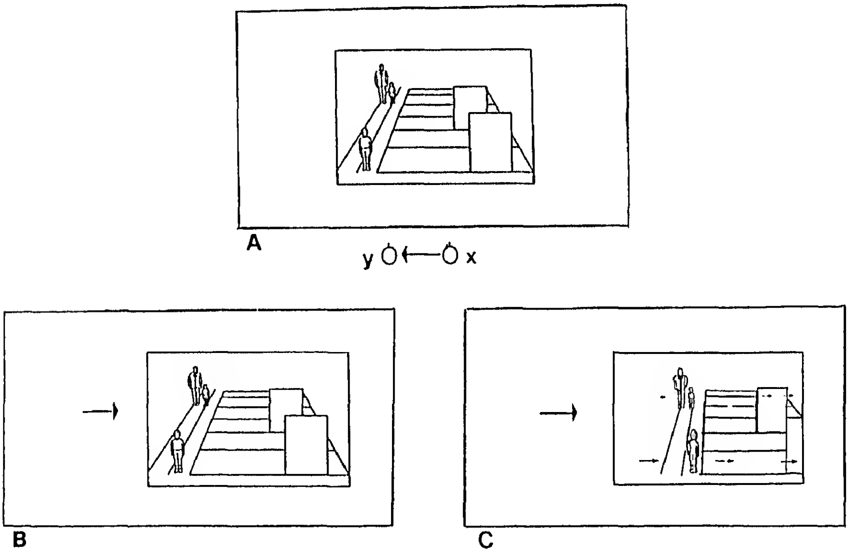
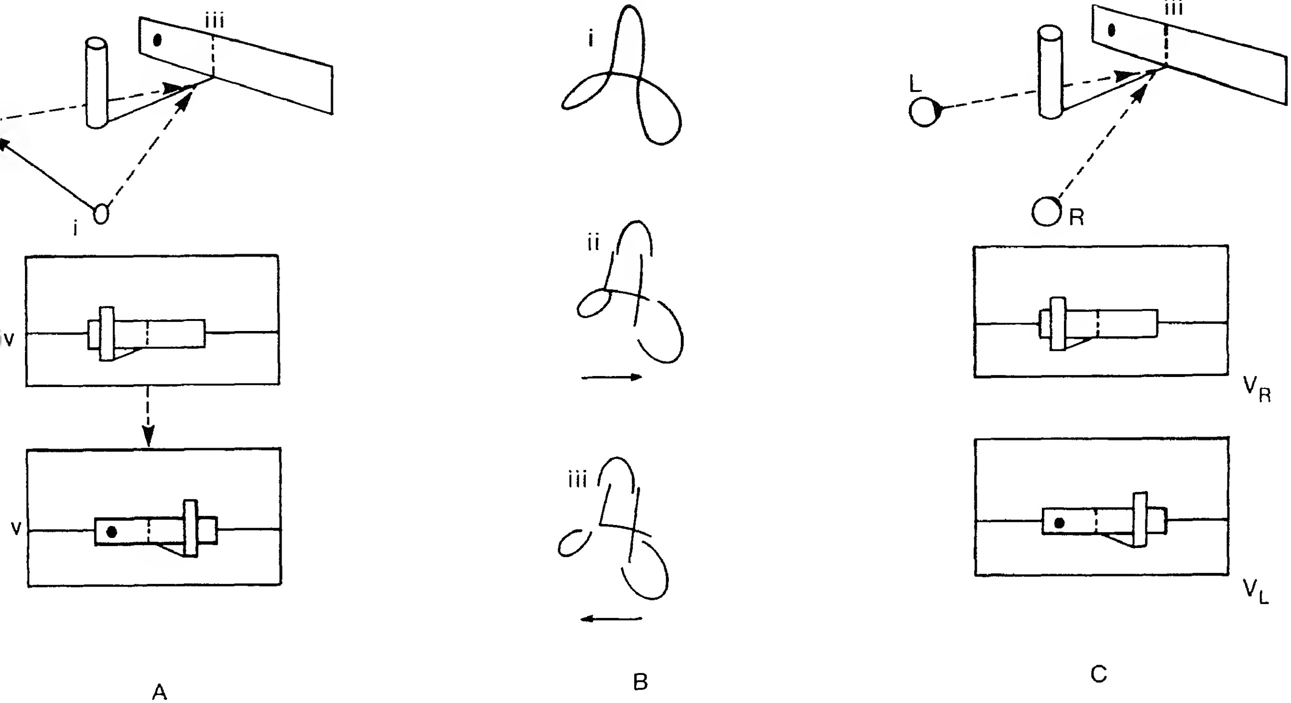


FIGURE 13 As Leonardo knew, pictures are not surrogates for a moving viewer. In A, a viewer moves from x to y. If the display is a picture, all parts are displaced equally in the field of view B, but if it is a window, objects at different distances undergo different parallax C. Those who take the invariants of the moving stimulus array to be fundamental to perception (see Figure 11) have yet to explain how it is that we perceive pictures.

The precision and ease with which we can study viewers' responses to moving patterns, and to other stimuli that change over time, depends of course on the equipment with which such stimuli can be produced and presented. Until the early 1950s only simple mechanical and electrical devices were in general use. Since then the dissemination of relatively cheap 16-millimeter motion picture cameras capable of producing controlled motion through animation, the advent of even cheaper and more convenient video equipment, and, above all, the availability of computer-generated displays, have progressively revolutionized the study of patterns that change with time. We are now in the midst of an explosion of research on the topic, done as much by computer scientists, physicists, and neurophysiologists as by perceptual psychologists.

Even the earlier and more primitive apparatus contributed a wide array of facts, some of which have been neglected in the recent interdisciplinary renaissance. Much of the earlier research was not directly addressed to questions of object perception but was intended instead to explore basic processes, e.g., the study of the time constants of the visual system's responses to flicker (Kelly, 1961), or the study of the conditions that yield *apparent movement* with successive simple static stimuli (Braddick, 1980; Kolers, 1972; Korte, 1915; Morgan, 1980). Some of the facts obtained in such research address the question of whether (and how well) our nervous systems respond to the stimulus changes that carry information about depth and motion (Figure 14). We know, for example, that our visual systems are extremely sensitive to motion parallax: even a very slight difference in distance between two aligned or nearby rods (Berry, 1948) and a very small head movement on the part of the viewer will provide a displacement in the retinal image that should be detectable (Helmholtz, 1866; Wheatstone, 1839). If two objects at different distances happen to line up from a particular view, therefore, and good continuation then provides a misperception of the object (as in Figures 12Ai [see p. 266], 14Bi), even a slight head movement should provide a detectable break in the good continuation.

Moreover, the two-dimensional shadows or projections of irregular spatial arrangements of rods, or of dots distributed in space, or of unfamiliar objects (Figures 15A, B, C, respectively), lacking other depth cues so that they are perceived as flat arrangements when stationary, are perceived as three-dimensional layouts when they are set into motion. Even more than the static Gestalt demonstrations, these phenomena seem difficult to explain as the use of a lookup table, learned by association, that the viewer can consult to determine the meaning of some previously encountered set of sensory events: How plausible is it that the viewer has encountered the particular pattern of moving randomly arrayed dots, shown in Figure 15B, so often that by familiarity it has become a recognizable tridimensional arrangement?



ɜURE 14 Motion parallax, binocular parallax, and their effects on adventitious alignments. At A, as the viewer moves from i to ii, :h gaze fixed on iii, the view changes from iv to v. At B, if the "4" happens to be adventitiously in perfect alignment with the ends the open loops in the background, a slight head movement to the right ii or left iii will provide a misalignment; this presumably should ːke good continuation (Figure 12A) inoperable except in static pictures. In C, without head movements, parallax is provided by the ɔ eyes' views. (R, L are right and left eyes; V_R, V_L are their views.)

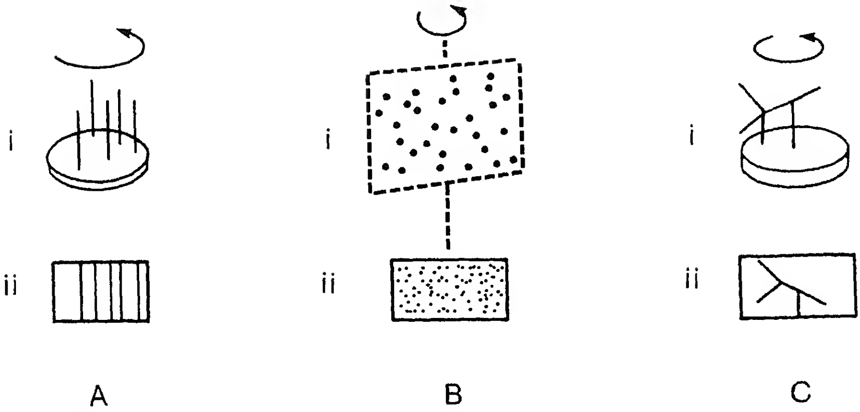


FIGURE 15 Structure through motion: Precomputer methods of studying motion perception. Unfamiliar patterns that look quite flat when static appear tridimensional when put in motion, encouraging the formulation that we perceive the rigid (or invariant) object that would provide the changing stimulus pattern. Some well-studied older examples are illustrated. A: The shadows of rods on a rotating turntable i, or the rods themselves, are viewed through an aperture that occludes their ends (Metzger, 1934; White and Mueser, 1960). B: The shadow of a set of dots on a moving glass plane i is projected on a screen ii (Gibson and Gibson, 1957). Such displays were initially the easiest to program and study in computer-generated form (Green, 1961). C: Simple unfamiliar wire forms mounted on a turntable provide a ''kinetic depth effect'' (Wallach and O'Connell, 1953).

It seems far more plausible that the phenomenon is the expression of a perceptual rule.

We have seen that we can in fact use relative displacement to discern spatial structure. But that still leaves open the question of what the rule is by which we fit three-dimensional space to the two-dimensional but moving stimulus pattern. As we have seen, the simplest and most general solution is that we extract that invariant object or layout that will fit the moving stimulation (Gibson, 1979; Johansson, 1980). This rule would account for the perception of rigid objects and surfaces without additional rules or constraints. Moreover, it includes the perception of motion pictures, and the phenomena represented in Figures 13 through 15, under the same general explanation.

As computers have made it easier to generate pictures of points moving in space, and as more research is done with such patterns, the point first made by the Gestalt demonstrations—that perception is governed by rules rather than lookup tables—has taken hold. And although Helmholtz and the earlier psychologists to whom perception is the result of learning often talked of what amounts to perceptual rules, no formal account has been offered of specific mechanisms or principles by which such perceptual learning might occur. However, once a rule is explicated with precision, it becomes relatively easy to imagine neural circuitry that might underlie its working. There is therefore added incentive today to take a ''nativist'' stance—to propose explanations of perception that depend on innate prewiring rather than on learning processes.

As we will see, the design of explanations in terms of such neural circuitry is a very active enterprise today, particularly in the field of computer science. But if that effort is to apply to human perception, it must start with perceptual rules that indeed are used in the human perceptual process. We still must decide what those rules are.

The perceptual rule that is most readily and explicitly defined in physical terms is the currently popular rigidity principle. In fact, however, *the strong forms of the rigidity principle will not work, for the perception either of objects in space or of their representations*. Evidence that decisively refutes the strong form of the principle includes findings obtained many years ago, although the implications of these facts have not been adequately taken into account in most recent discussions. The same facts make the other overarching principles, as they are presently conceived, equally unworkable. We survey some of that evidence next.

Some of the points that follow have recently been made as well by Braunstein (1983), by Gillam (1972), and by Schwartz and Sperling (1983).

Why the Strong Forms of Various General Perceptual Principles Must Be Rejected Although the overall case against these general rules cannot be reviewed here in detail, the strongest argument is simple and sufficient: Even when rigid moving shapes are in full view, we do not necessarily see them. In some cases we perceive instead quite different shapes undergoing nonrigid deformation.

This has been known in a general way at least since 1922 (e.g., von Hornbostel found that a real, rotating wire cube reverses perspective even though it must then appear to stretch and bend), and a remarkably robust illusion known as the Ames ''window'' has been widely disseminated since 1951: A trapezoid (often with shadows painted on it to ''suggest'' the perspective view of a window) rotates continuously in one direction (e.g., arrow vii as seen from above in mirror, M) either clockwise or counterclockwise, in full view, as shown in Figure 16A. It is not seen as such. Instead, it is perceived as oscillating (arrow viii), reversing direction twice each cycle so that the larger end (i) (or iv in the mirror) always appears the nearer. It is as though a process of unconscious inference were at work, assigning depth on the basis of the static depth cue of linear perspective (Figure 2, see p. 251) and inferring direction of movement from relative

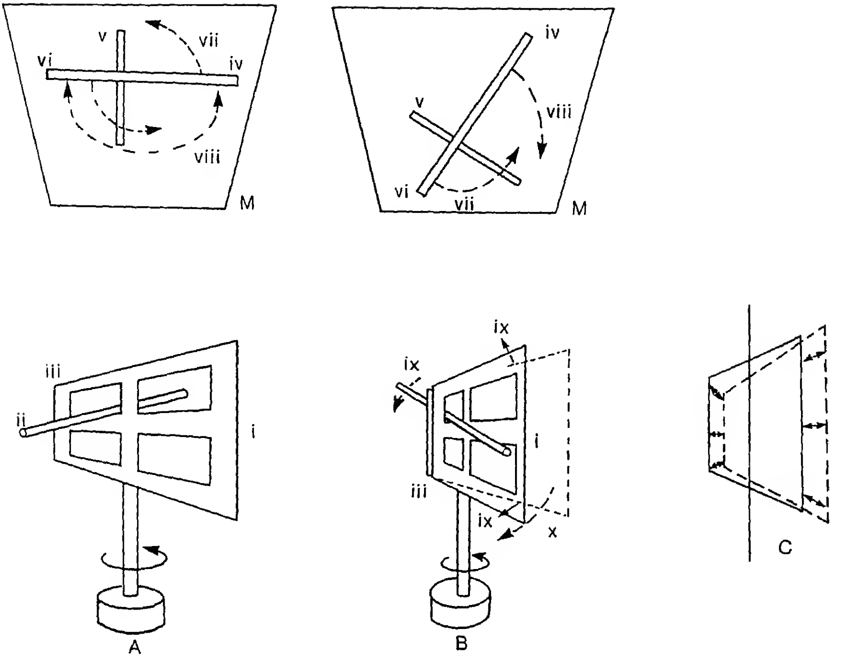


FIGURE 16 A classic illusion with a moving object. In A, a flat trapezoid, with markings painted on its surface to "suggest" depth, is seen from in front, with i and iii equidistant from the viewer, and from above in the mirror M. At B, the trapezoid has rotated so that the small edge iii is nearer the viewer. Ames (1951) found that though rotating continuously (arrow vii in the top view) it appears to oscillate back and forth (arrow viii); see text. Although the rod, ii, is rigidly fixed to the trapezoid, it is correctly seen to rotate, passing *through* the substance of the trapezoid! The trapezoid cannot both appear to oscillate and yet remain rigid in appearance. The solid and dotted outlines in C are its shape as presented to the eye when edge iii or i is respectively the nearer. If seen to oscillate, the trapezoid must also appear to deform between these shapes, as shown by the arrows, although this nonrigidity is not normally very noticeable.

depth. I hasten to add that although it is widely offered (e.g., Ames, 1951; Gibson, 1979; Graham, 1963; Hochberg, 1978b) there is no experimental support for such an explanation of the phenomenon; indeed, there are features of the changing retinal image that might be direct, if misleading, bases of the illusory response (Braunstein, 1976; Hochberg, 1984b). (For example, even when the larger end swings away from the viewer, as shown by arrow vii in Figure 16B, a vector of expansion, ix, will generally be provided as the large end swings in toward the axis of rotation, and expansion is normally a correlate of approach; cf. Figure 11C, p. 264.)

This illusion is extremely strong and difficult to overcome even when the viewer is confronting a real object (although in that case, one eye must be kept covered at close distances; at really close distances the true shape and movement may be seen monocularly as well). When the viewer confronts a moving picture, rather than the object itself, the illusion is almost irresistible. The virtual shape that then fits the perceived illusory movement to the changing pattern of light at the eye must then be nonrigid (Figure 16C), and the perceived path must follow a complex and changing radius. Moreoever, the illusory 180-degree oscillations of the trapezoid are perceived even if a rod is rigidly affixed to the trapezoid, as in Figure 16A and B; the rod does not appear rigidly fixed, but pursues its 360-degree rotation, apparently passing *through* the trapezoid like a phantom when the trapezoid reverses its apparent course. (This is true even if the viewer is simultaneously shown the setup from above in a mirror; the rod and trapezoid are seen to rotate 360 degrees as a rigid unit in the mirror, and, at the same time, to move in separate parts in direct view.)

Thus, a truly rigid and invariant object, moving in a simple and invariant orbit, is not perceived, and instead a nonrigidly deforming and quite illusory object is seen, moving in a complex and variable path.

This phenomenon, although widely popularized since 1951, has been virtually ignored by those who propose that our perceptions are determined by invariance, rigidity, or simplicity principles. I can find only brief mention of the phenomenon (Gibson, 1979), claiming that it occurs only when the motion-provided information is below threshold, and that then the illusion rests on unconscious inference. This highlights the question of thresholds, which must surely be considered before we can say that any of the motion-produced information discussed in connection with Figure 11 (see p. 264) provides anything useful to the viewer, and which has yet to be addressed in any systematic evaluation of the direct theory (Cutting, 1983; Hochberg, 1982). Moreover, by invoking unconscious inference, this way of dealing with the phenomenon spoils the direct theory's claim to parsimony. But in any case, that answer is wrong. Even when the changes provided by the moving object are clearly above the detection threshold and the illusion is therefore accompanied by clearly perceived nonrigidities, the latter is what we see, and not the veridical rigid motion (Hochberg, 1984b; Hochberg et al., 1984).

Subtle arguments are not needed, however: Given the lessons of Figure 16, we can readily devise new illusions in which rigid simply moving objects, freely viewed (with monocular vision), are seen to bend and deform nonrigidly, as in Figure 17. Motion is not enough to ensure veridicality, therefore, and what is perceived may be perceived *against* any effects that simplicity, invariance, and rigidity principles might exert.

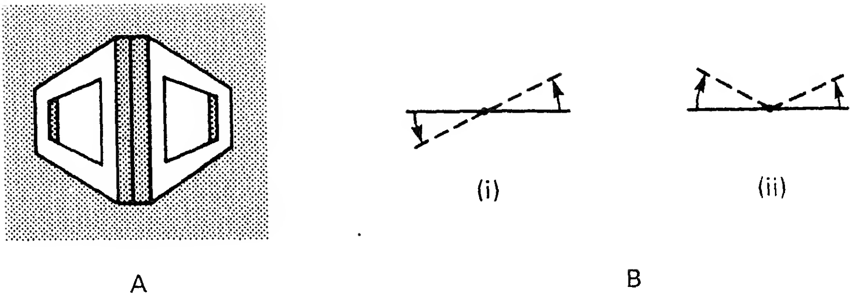


FIGURE 17 Apparent bending in a rigid, moving object. A flat, rigid octagonal cutout, with markings painted on its surface to "suggest" depth, is shown in front view at A and from above at B. To monocular vision, when it moves as shown in Bi, it tends to appear instead to hinge in the middle, and to "flap" away from the viewer as shown in Bii. Similar but less compelling effects occur without the markings, and with oval shapes as well (Hochberg and Spiron, 1985).

COMPUTERS AND PERCEPTUAL PSYCHOLOGY

The microelectrode was a major technological watershed, and its effects were quickly manifest. The introduction of the computer has had far greater effects, but they have been more diffuse, are slower in being realized, and are still growing, as computer science and technology change.

There are six main ways in which the computer has affected perceptual psychology; although these ways are closely intertwined, they are also very different, and it is important to separate them if one is to understand the relationship between the two disciplines.

The first two uses are contributions that computers now offer every branch of science: obtaining and analyzing data, and modeling theories and explanations.

Obtaining Data

The computer has of course radically changed the methodology of measurement and analysis. For example, the direction and changes in the subject's gaze can be monitored and even used to control the display that confronts the eye (McConkie and Rayner, 1975), permitting the detailed study of how the integration of successive glances occurs in the process of reading text and perceiving pictures. Such research would simply have been impossible without high-speed and powerful computers; we will see that the problem to which this method is addressed is of central importance. By handling large quantities of numbers and rapidly executing operations that were once prohibitively time-consuming and expensive, the computer makes available data that were in essence unobtainable. This is true both of physiological data and of judgments intended to tap perceptual experience. Thus, physiological signals that are normally far too weak to be distinguished from accompanying bioelectrical background noise can be adumbrated by computer methods that cumulate them over many occurrences, making it possible to measure the electrical potentials (Donchin et al., 1978; Sutton et al., 1965) and magnetic fields (Kaufman and Williamson, 1982; Reite and Zimmerman, 1978) at the scalp. Such averaged *transients* reflect neural responses that the brain makes to sensory stimulation and that accompany perceptual processing.

Modeling Theories and Explanations

The second use of the computer, common to all science, is to model theoretical proposals and explanations for which it would otherwise be impossible or too laborious to say whether and how they would work. Whether some hypothetical neural network would respond as designed (Hebb, 1949; Marr, 1982; Rashevsky, 1948; Rosenblatt, 1962), whether a particularly defined set of flow patterns (like Figure 16C) would specify uniquely a set of surface forms in the world (Ullman, 1979), whether a particular history of strengthened associations would even theoretically result in perceptual learning (Hebb, 1949; Minsky and Papert, 1969; Rosenblatt, 1962)—these are questions that cannot be answered simply by considering them in verbal form but that can often be answered once the functions are stated specifically enough to run as a computer program.

Embodying Perceptual Functions

A second branch of computer science aims at embodying perceptual functions, similar in effect to those of humans, in computer hardware and software. We must distinguish two distinct purposes that guide this enterprise. One is to design and provide devices that can serve instead of humans. Some of these functions are readily achieved (the sensors that open supermarket doors, the bar-code scanner that identifies and prices items at the checkout counters), and some are probably unachievable in the foreseeable future (e.g., machines that respond to or translate free and normal human discourse); but in general there is no compulsion to serve each function in the same ways that humans do. Human perceptual functions here serve only as ''existence'' proofs that assure the computer scientist that at least one way of solving the problem exists and is embodied in human neuroanatomy.

Once we start to consider the means by which modern electronic computers might perform such tasks, however, we develop new ways of thinking about how the human nervous system performs its perceptual tasks. The computer then serves as an analogy or even a model for the study of human perceptual processes. That may turn out to be the most important relationship of all between computer science and perceptual psychology, and we consider that next.

The Computer as an Analogy to Perception

Perhaps the greatest effect of the computer has been its influence as an analogy: Inherently vulnerable to entrapment in the mind-body problem of philosophers and metaphysicists, and self-conscious about the need to be scientific, psychology is always tempted to confine its attentions to variables that are conceived and measured in physical terms. Indeed, almost since J.B. Watson's behaviorist manifesto in 1913, physical measurement and physical (or at least physiological) conceptions have enjoyed intellectual hegemony in this country.

There was of course continuous opposition, both on scientific and metaphysical grounds, and the field of perceptual psychology by its very subject matter was less constrained by behaviorism than other fields of psychology, but for that very reason it was almost abandoned as a discipline for some two decades. It was not until the late 1950s that what can only be called ''mental'' conceptions and measures once again became scientifically respectable to the rank and file of the profession. I am convinced that the main factor in this change was the obvious fact that computer programs are in principle transportable to very different physical machines. They can therefore be analyzed and discussed in abstract functional terms without reference to the specific hardware in which they must be embodied to perform. Familiarity with computer functions, terminology, and flow charts made it possible to describe what the mind might be doing in a way that could, in principle, be instantiated in a program and then embodied in a machine (Miller et al., 1960; Rosenblatt, 1962; Selfridge, 1959).

Something like this had already been done repeatedly, long before computers were developed, from Descartes' design in 1650 of a hydraulic model underlying neural function, to Tolman's analysis of purposive learning by a ''schematic sowbug'' in 1938; but there was never any real likelihood that the analyses might be put to the test by building the machines. The general-purpose computer and transportable programs have made the point much more powerfully.

The language of cognitive psychology is now very close to the language of computer science. There is usually no guarantee that any given flow chart with which the cognitive psychologist offers to explain some phenomenon can in fact be translated into an executable program, but if not, it is inadequate because it is vague or inconsistent and not because it is mentalistic.

Computer Science in Perceptual Psychology Research

The attempt to design machines that embody human perceptual functions (or to design programs that model such machines) rests on the belief that only in this way can we be sure that we have achieved a scientific understanding of those functions. This is an old belief and undertaking, but the advent of the modern computer makes the venture seem more plausible. Given its purpose, this undertaking must start with scientific empirical knowledge of how humans perceive. That is of course precisely what the task of perceptual psychology has been. In consequence, the two disciplines now overlap greatly, and an increasing amount of perceptual (and cognitive) psychology is currently being done in computer science departments.

This is a very promising development, and some of the work has received wide attention as a ''breakthrough''; but it would be unwise to overestimate what has been accomplished at this early stage (see Braunstein, 1983; Haber, 1983). The approach ensures precise modeling of theories but does not by itself provide either new theories or new facts about human perception. The point is worth spelling out in a brief examination of the field.

Because it is far easier to make initial progress at formulating specific models of direct neural response to stimulus information than at formulating specific models of central processes of learning and inference, most of the work in this field has concentrated on the former (see Haber, 1983). As a first stage, any perceiving machine must be able to separate objects from their cluttered surroundings; this problem is very difficult to deal with in still pictures (Oately, 1978; Roberts, 1965). We have seen above that the problem is mathematically less refractory, given the multiple views provided by motion parallax and binocular parallax (see Figure 14A and C on p. 271) in that fewer constraints are needed to specify the three-dimensional layout that would produce the stimulation at the eye. It is understandable, therefore, that computer scientists have recently turned to models of binocular stereopsis (Marr, 1982; Marr and Poggio, 1979) and of the perception of structure through motion (Marr, 1982; Ullman, 1979).

These ''computational'' models are totally within the mainstream of perceptual psychology (although that is not always clear from their presentation, nor from their reception). For example, the computational model of binocular stereopsis devised and tested by Marr and his colleagues was a relatively slight variation of a detailed theory published in 1970 by Sperling, a psychologist; Sperling's theory is itself well within that class of psychologists' explanations of stereopsis (see Kaufman, 1974) that have taken Johannes Kepler's (1611) geometrical analysis of the binocular lines of sight that obtain when viewing objects in space as the model of an internal "binocular neural field" that merely reflects that geometry. And Ullman's computational analysis of the information that moving stimuli give about their layout in space takes its place within the long tradition of such analyses and research. Neither of these computational perceptual theories can claim to be more than partial accounts of the phenomena in the domains they address. For example, even with unimpeded binocular parallax, we perceive the concave mold of a human face lit from below as a convex face lit from above; even with unimpeded motion information, as we have seen (Figures 16–17), at least some rigid objects are perceived as nonrigid and in wrong slant and motion. These computational theories do not differ from other attempts at sensory explanation of object perception in their inability to deal with such problems. They differ only in that they are restricted to models that can be successfully run as computer programs, and that is not necessarily an unalloyed virtue.

Although computer simulation and "computer perception" have received considerable praise in recent years, there are grounds for criticism as well. The need to devise perceiving machines that work as humans do is certainly not a valid economic argument—one can usually find far more direct means of performing specific tasks. Nor is computability a necessary criterion for assessing any theory, regardless of how desirable that quality may be (and despite the stress on simulation studies currently evident in many quarters).

But these arguments are moot. Regardless of the intrinsic merits of computer simulation and of the quest for perceiving machines, and without appeals to metatheory or philosophy of science, there remains a present and growing need to develop theories of human perception to the point that they can be embodied in computer programs. That is the relationship between the computer and perceptual psychology that we consider next.

Why Models of Human Perception Are Needed

Computers communicate to their human users through pictures as well as through words and numbers. But more than that, they are increasingly used specifically to *generate* pictures: as interfaces between the viewer and some part of the world that would otherwise be difficult or impossible to see; as means of visualizing designs of buildings, machine parts, molecules, chromosomes, or cellular processes; as substitutes for human artists and animators in creating graphic displays for advertising and entertainment; as simulators in flight training. The use of such devices is already great, and growing rapidly. In many cases, the pictures (or pictorial sequences) that are displayed were not themselves programmed and were never previewed in any sense, but are generated in response to the question that the user asks or in reaction to something that the user does. An example of the former would be what an architectural layout or a machine part will look like from some chosen location, or what the state of a flow chart would be under some specified conditions; an example of the latter would be a low-altitude flight simulator display, which depends, of course, both on the terrain being simulated and on the individual pilot's actions.

Without a human editor intervening between computer and user, there is some unknown likelihood that the viewer will be shown misleading or incomprehensible pictures. Where that likelihood must be minimized, the computer must avoid certain classes of pictures, or must be prepared to enrich or enhance those pictures. This means that we must be able to specify, in terms acceptable to a computer, how humans will perceive a picture or a sequence of pictures.

The study of the rules of representation is now a vigorous and growing field in which perceptual research finds practical application (Cutting and Millard, 1984; Haber and Wilkinson, 1982; Stevens, 1983; Todd and Mingola, 1983). Although this task shares much of what we must learn through exploring analogies between computers and human perception, it is also significantly different. It cannot ignore as mere embarrassments the cases in which we misperceive, the exceptions to proposed generalizations—indeed, it is just those cases that must be the focus of inquiry. And that is fortunate for psychology, because those are the cases that test the generality of any perceptual theory.

A superficial answer to the question of how we can ensure comprehensible pictures is to increase the fidelity of the surrogate—i.e., to make the light to the eye more like that provided by the object or scene that is being represented. That means improving the resolution and the color balance, avoiding distortions, etc. Indeed, if other things are equal, an improvement in these engineering factors will usually improve picture comprehensibility. But we have seen that even perfect fidelity—i.e., the moving object itself—may result in misperceptions (Figures 16, 17). The constraints on mental structure—on the structure of perceived objects—are not the same as the normal constraints on physical objects, and we must know the former as well as the latter if we are to be able to predict how pictures are perceived, even with the best picture quality possible.

There are practical limits, moreover, to the pictorial information that we can count on. One can see detail in a closeup, or an entire object or scene in a long shot, but not both. Picture quality is limited, and the techniques that motion picture and television filmmakers have developed to cope with those limits—surveying or scanning an object or scene by successive partial views or closeups—require the viewer to go beyond the momentary sensory input, and to enter and store the successive partial views in some mental or perceptual structure of the object.

Perceptual structure refers to the relationships *within* what one perceives (Garner et al., 1956; Hochberg, 1956), that is, to information about the object that can be retrieved from the viewer—for example, that the sides of a cube look equal and parallel, that the vertical edge at *1* in Figure 12E (see p. 266) looks farther than the horizontal edge it intersects when the vertical edge looks the nearer at 2. To the degree that perceptual structure reflects the structure of the physical stimulus, physical analyses of optical information will serve to model the perceptual process; obviously, as long as we stipulate that some object, say, a wire cube, is perceived correctly, the layout of the physical cube itself must serve to predict the relative apparent nearnesses of its parts. The simplicity of this task is of course what makes the more extreme direct theories so attractive. To the degree that perceptual structure reflects known (or hypothetical) neurological structure, however, the latter must also modify any attempts to relate what the viewer perceives, on the one hand, to the optical structure of the object or scene that confronts the perceiver, on the other. Thus, what we know about the distribution of acuity over the retina or what we think we know about spatial frequency channels must be used in attempting to predict the effects of the information that could otherwise be provided by the optical structure. Computer models of perception can incorporate both kinds of structure with very little input from psychological research.

To the degree that perceptual structure reflects none of these—that is, to the degree that it expresses what we may call mental structure—perceptual research must provide the facts that are needed for any theories, whether or not those theories are embodied in computer models. Such facts are obtainable but sparse; for this reason, computer science has as yet very little to say about the modeling of mental structure. Some terms have been offered (e.g., Minsky's ''frames'' [1975], roughly equivalent to an expectation or a schema), but terms or even models are not needed here so much as facts, and more attention paid to what facts we do have. We next consider very briefly the current state of research on mental structure in real and represented objects.

MENTAL STRUCTURE IN OBJECT PERCEPTION AND REPRESENTATION

Where perception can be predicted from the pattern of stimulation that falls on the sense organs, we are free to argue that the stimulus pattern itself (as transformed and limited by the sensory system) determines what we perceive. Of course, we are also free to hold that there are other factors at work as well. The fact is that we have a long history of demonstrations that sensory stimulus information is neither necessary nor sufficient to determine what we perceive.

One source of such evidence is provided by completion phenomena, examples of which are shown in Figures 18 and 19. These should not be dismissed as strained: In a normally cluttered world, such interrupted and fragmented shapes must be the rule rather than the exception. Nor can our perceptions of these shapes be profitably ascribed either to complex sensory structures (such as receptive fields and frequency channels) or to invariant stimulus information. The perception of a single object rather than of separate fragments often depends on the viewer's having specific knowledge of what that object normally looks like, and on being ready to perceive it. That is very much what Mill and Helmholtz meant by perception. A nice demonstration to which Dallenbach called attention in 1951 is shown in Figure 19A, in which few viewers can discern any clear object. After looking at Figure 19B, however, it is remarkably difficult *not* to see that same object when looking at Figure 19A.

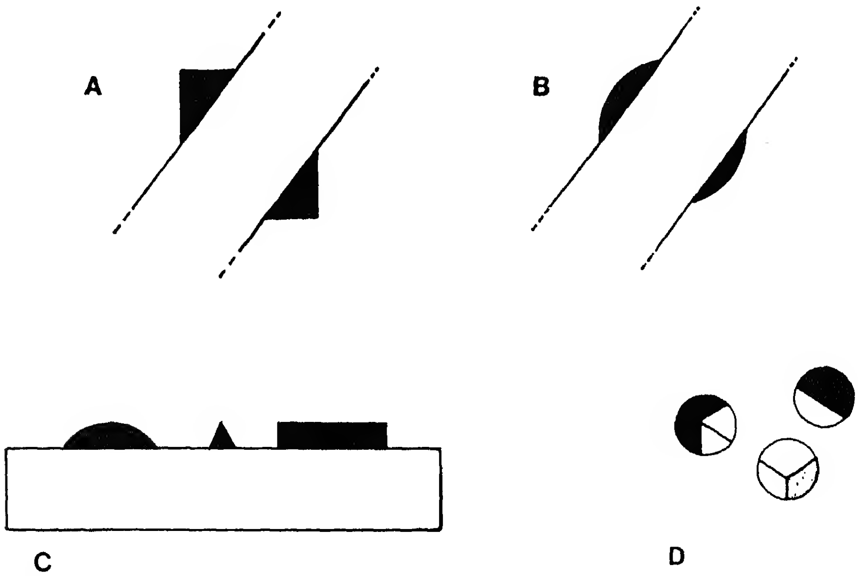


FIGURE 18 Completion phenomena. A selection of simple geometric shapes: at A, a square; at B, a circle; at C, a circle, triangle, and square; at D, a cube. In fact, the fragments that are shown do not by themselves define or specify any shape. It might be, for example, that C consists of the block letters CAT, partially occluded.

Completion phenomena have been known to psychologists for more than a century (and to artists, of course, for much longer). While the classical theory prevailed (see Figure 6, p. 254), such demonstrations were taken for granted and assumed to reveal the pattern of associations (or sensory expectations) that each viewer has learned from experience with any object. Although viewers would differ in their individual perceptual histories, the structure of their sensory expectations or associations should nevertheless reflect at least grossly the covariations or contingencies of the physical world—as filtered through their limited sensory systems. That is, mental structure should be predictable from measures of physical structure (e.g., Brunswik, 1956), once sensory limits are taken into account.

In some cases, mental structure does indeed seem to be at least approximately that of physical structure (e.g., the "constancies" described in connection with Figure 7), but we also have had dramatic counterexamples for the past 30 years (Figures 12F, 16A). Some psychologists still adhere to "unconscious inference" explanations today (Gregory, 1970; Rock, 1977), but such counterexamples make that proposal as it now stands an empty one. To mean anything at all, the premises of such supposed inferences must be investigated and not simply taken to be the same as the structure of the physical world. Moreover, as we have discussed at length, the last 30 years have also shown that much of perceptual structure may be given directly by complex neurophysiological circuitry; if that is at all true, such prewired perceptual structure must surely affect the nature and use of whatever mental structure does exist in addition. For example, for all we know at present the Ames trapezoid phenomenon (Figure 16) may result not from unconscious inference but from some direct sensory mechanism that provides a salient illusion only in certain conditions (see Figure 16C, p. 274; Hochberg, 1984b).

We need, therefore, to study mental structure and to measure its characteristics. The very topic has an aura of insubstantiality, until recently anathema to many psychologists eager to avoid subjectivity and mentalism. To study how a person perceives some object we must in one way or another ask him or her questions about that object—retrieve information from the subject about the object. In cases like the completion phenomena, we must ask the viewer questions about an object that is not in fact present and for which only that absent virtual object, and the few stimulus fragments actually shown to the viewer, can be confidently described in physical terms. That fact is a challenge but not an insuperable obstacle. There is actually a considerable body of research with a much more extreme experimental situation: Since Galton (1883) first undertook to study individual differences in mental imagery, methods have existed for studying how well individuals can retrieve information about objects for which *no* stimulus information whatsoever is present. Such "objective tests of imagery" (see Woodworth, 1938) have seen increasing use in recent years, but rather than being used to probe individual differences in imagery, most present research is directed to examining the nature of the imagery process itself (Kosslyn, 1980), a task that faces many of the same challenges as does the task of studying mental structure in perception. Whether such imagery studies have substantial implications for perception is unclear. We do not know whether imagery, studied with no stimuli present, is related in any simple way to the mental structure that is involved in the perception of partially present objects. That can only be answered by research on the process by which mental structure informs and accepts sensory information.

The need to fit fragments of sensory information into some mental structure is pervasive in normal perception. The perception of objects that are partially obscured in normally cluttered environments must often draw on a process of fitting fragmentary sensory information into a previously provided mental structure (Figure 19).

In addition, our perceptions of any scene or moderately large object must be assembed over time by means of successive glances, each of which provides only a partial view of the world. Finally, as objects are temporarily obscured by nearer ones (as viewer, object or both move through the world), we must be able to keep track of their motions even while they are out of sight, and to recognize them when they reappear. Both of these functions are drawn upon in our perceptions of real objects in the world and also in

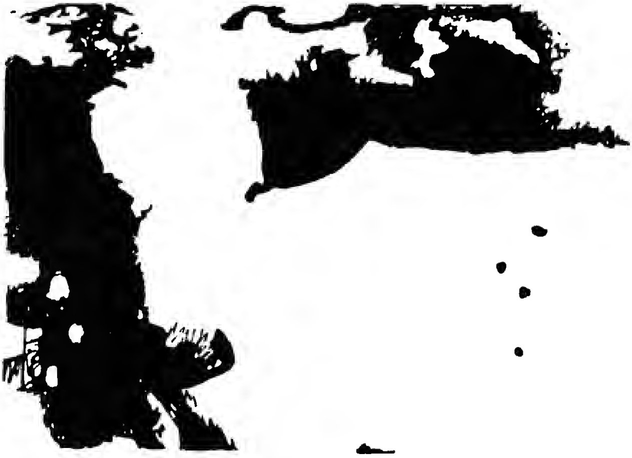


FIGURE 19A A completion figure. The mysterious object shown in this high contrast photograph is more readily apparent in Figure 19B on page 286 (reprinted, with permission, from Dallenbach, 1951).

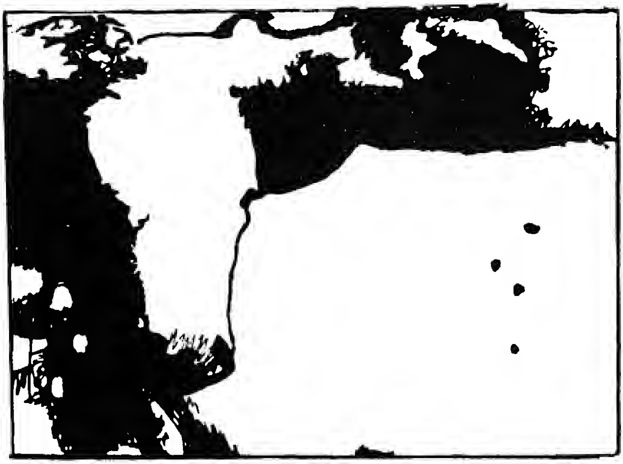


FIGURE 19B The same cow shown in 19A (reprinted, with permission, from Dallenbach, 1951). Once the object has been seen in Figure 19B, it is remarkably difficult to avoid seeing it in 19A as well.

film and video, as cameras cut from one scene to another (both successfully and unsuccessfully). And both functions suggest methods by which mental structure may be studied. We consider these in turn.

In the normal process of directing our gaze at different parts of some object, each glimpse offers detailed vision only in a small central part of the retina. The information gained by the successive fragmentary glimpses (as many as four per second) must therefore be integrated by some nonsensory process into a single perception. Similarly, in virtually all motion picture or video sequences, successive closeup views or shots each provide a partial view of some scene that may never be shown in its entirety (which would be a long shot) and may in fact not exist at all save in the mind's eye of the viewer. This is a kind of completion over time, of central importance to perceptual theory and application, that could not be studied at all until the last decade, when motion pictures and high-speed computer graphics became generally available as laboratory tools.

There has as yet been little more research on this aspect of object perception than to show that such research is possible. The row of circles in Figure 20 represents a sequence of successive views that simulate a stationary circular aperture through which the individual corners of some object that is being moved about behind the screen—in this case, a cross—are visible. If the motions of the corners were themselves visible, the viewer could construct the entire object behind the screen in his mind's eye, detecting, for example, that a specific arm of the cross has been skipped

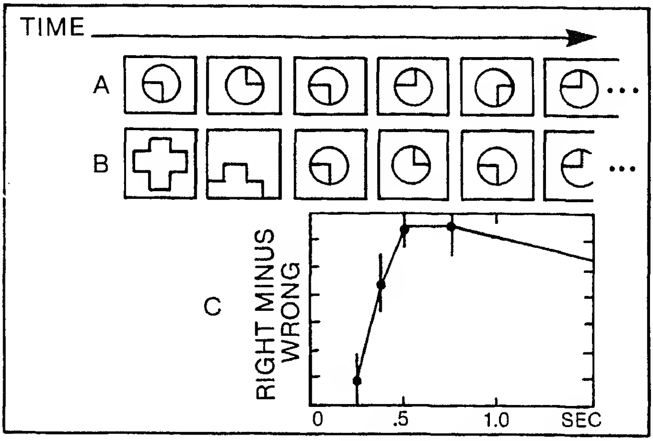


FIGURE 20 Other completion figures. A sequence of right angles, presented at rates of 333 to 2000 msc per view, is shown at A. Subjects who are shown two such sequences, which may or may not differ in one or more views near the middle of the sequence, cannot tell better than chance (the baseline in C) whether the sequences are the same or different, because each sequence of views, considered as independent items, far exceeds their memory span. In each pair of sequences, at least one sequence is in fact a systematic succession of closeups of the corners of a cross. If each sequence is introduced by a long shot and a medium shot, as at B, which establishes the overall object and the starting point of the sequence, then each view of the unchanged sequence takes its place in turn within the structure that the viewer has in mind, whereas the altered sequence does not, and the difference between the two sequences (which are no longer strings of independent events) becomes evident, within the time limits indicated at C.

within the sequence. If the motions are not visible, as they are not in this experiment, then the sequence of static views is indecipherable and in fact cannot be kept in mind; if a long shot of the object is presented first, however (as in row B), providing a mental structure within which the successive views can take their place, the subject can again perceive the object that is moving behind the aperture (Hochberg, 1978a). It is the mental structure of the object that makes the stimulus sequence comprehensible. Given the long shot and the structure it provides, two sequences that are different are perceived as such; without the structure, the viewer cannot distinguish one sequence from another.

When a pedestrian you are watching is lost from view while he passes behind a parked truck, or while you divert your gaze to the traffic light, you must still be able to tell approximately when he will return to view from behind the truck, or where he will have gotten to when you look back from the traffic light. Such predictive functions, for which we can surely find ample evolutionary demands, imply that something that corresponds to motion through space occurs in the mind's eye of the viewer. The filmmaker or graphics programmer who cuts away from one event to another, and then returns to the first one, must make some assumptions about how well the viewer keeps track of any motion that is going on in the first event. The following research shows that discussing such mental motion is more than just a poetic metaphor.

Shepard and his colleagues had shown in a wide range of experiments that the time subjects need to judge whether two objects are the same or different (Figure 21A) is proportional to the angle between them, as though one object were being mentally rotated at some constant rate to bring it to the same orientation as the other in making the comparison (Shepard and Cooper, 1982; Shepard and Metzler, 1971). Using that paradigm, Cooper (1976) first determined each subject's characteristic ''mental rotation'' rate, ω, and then, after having had subjects memorize the figures, displayed the comparison figure at some variable angle (ϕ) and delayed after a starting signal by a variable interval (t). She found that if the product of $\omega \times (t) = (\phi)$, judgment times no longer increased with angle (ϕ): they were now independent of the angle between the two objects being compared (Figure 21B). The results are what one would expect if the object had in fact been rotated at angular velocity $\omega \times (t)$ between presentations i and ii, and if both objects had come to the same orientation by the time the comparison was called for.

Given these findings, ''mental rotation'' seems more than a metaphor that summarizes the fact that judgment time is a function of angle (ϕ) in Figure 21A. It implies a usable and consistent relationship between time and distance in a mental structure that cannot be attributed to physical stimulus information.

A third and quite different paradigm, which appears in a recent technical report by Cooper (1984) on work in progress, may tell us something more general about the form in which perceived objects are manipulated and stored and may also eventually provide a tool with which to compare how well different methods of representation accord with the ways in which objects are perceived and remembered. Subjects had been given two orthographic projections of an object (a and b in Figure 22) and were to judge whether a third orthographic projection (c) was of the same or of a different object. No isometric projections (e.g., c) of any objects were shown to the subjects at this time. Subsequently the subjects were shown a set of isometric projections (e.g., c, f), some of which represented the objects used in the previous tasks and some of which did not. Subjects tended to report that they had seen the former before, even though no isometric pictures at all had been shown. Although this research is still in progress, and various

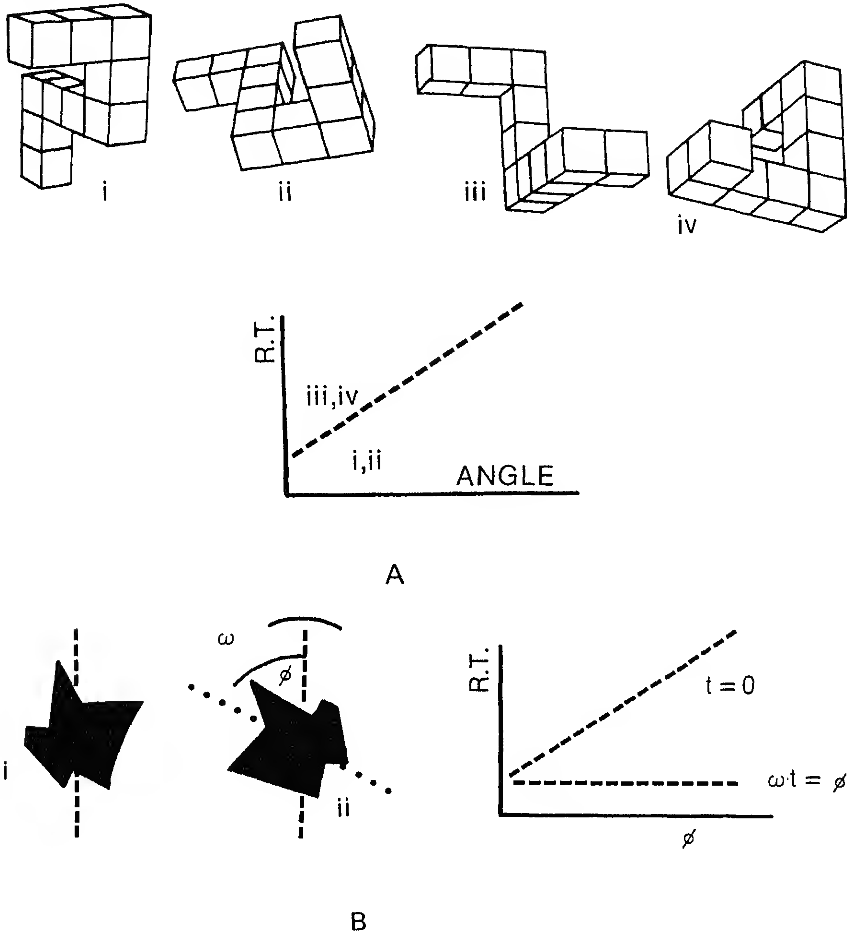


FIGURE 21 Mental rotation. A: Given two objects at different orientations, the time that it takes to judge whether they are the same or different is a function of the angle between their orientations, whether in the picture plane i, ii or in depth iii, iv. It is as though the subject must rotate one object into the orientation of the other before the two can be compared (Shepard and Metzler, 1971). B: If the two shapes to be compared i, ii are presented simultaneously (i.e., separated in time by an interval $t = 0.0$) their reaction time R.T. increases with angle, ϕ, between their orientations, as above. But if the comparison figure is presented after an interval $t = \phi/\omega$, where ω is the subject's characteristic rotation rate (obtained from the slope of the function at $t = 0.0$), then the R.T. does not increase with increasing angle ϕ (Cooper, 1976). This is just what one would mean by saying that the subjects had rotated the object before making the comparison.

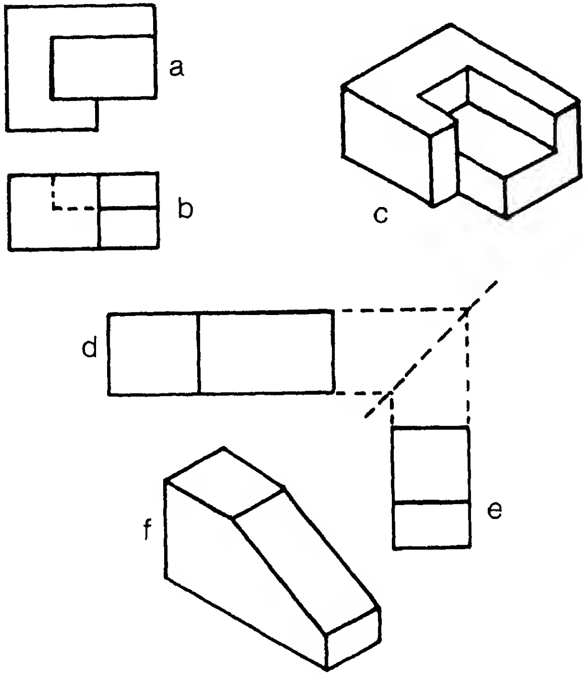


FIGURE 22 The structure of perceived objects—orthographic and isometric projections. Two different pictorial systems are shown here: Pictures a and b and pictures d and e are orthographic projections of objects whose respective isometric projections are c and f. Isometric projections are easier to grasp, at least for these objects. Cooper (1984) presents preliminary evidence that even when subjects have been presented only with orthographic projections of objects, they tend to report later that they have seen isometric projections of those objects.

controls are needed, the preliminary results will, I feel certain, survive the necessary replication and controls: Orthographic and isometric projections can both specify the form of a three-dimensional object, but the isometric projections are in some sense closer to the way in which we extract and store the information—closer to the mental structure involved in perceiving and comparing the objects. Although I know of no research to the point, it hardly needs an experiment to discover that isometric pictures are more rapidly and accurately comprehended than orthographic ones. What experiments can do is give us a better understanding of why that is so, and of the sense in which the isometric picture is more like the structure that underlies our perception of the object.

These three research procedures that I have described in connection with Figures 20 through 22 are interesting more as examples of a field of experimental and quantitative inquiry than as demonstrations that mental processes can be studied and are in that sense real; the latter is not a new conception. It has repeatedly come into and gone out of scientific fashion, and merely showing that mental structure "exists," in some sense, will not add much to its history. Fortunately, this time there are vested interests in obtaining and systematizing the knowledge, and technical facilities for doing so, that should keep research and theory centered on these problems of object perception and representation for some time to come.

REFERENCES

Ames, A.

1951 Visual perception and the rotating trapezoidal window. *Psychological Monographs*, No. 324.

Attneave, F.

1954 Some informational aspects of visual perception. *Psychological Review* 61:183–193.

1959 *Applications of Information Theory to Psychology*. New York: Holt, Rinehart and Winston.

Barlow, H., Blakemore, C., and Pettigrew, J.

1967 The neural mechanism of binocular depth discrimination. *Journal of Physiology* 193:327–342.

Bekesy, G., von

1960 Neural inhibitory units of eye and skin. Quantitative description of contrast phenomena. *Journal of the Optical Society of America* 50:1060–1070.

Berry, R.N.

1948 Quantitative relations among vernier, real depth, and stereoscopic acuities. *Journal of Experimental Psychology* 38:708–721.

Blakemore, C., and Campbell, F.W.

1969 On the existence of neurons in the human visual system selectively sensitive to the orientation and size of retinal images. *Journal of Physiology* 203:237–260.

Braddick, O.J.

1980 Low-level and high-level processes in apparent motion. In H.C. Longuet-Higgins and N.S. Sutherland, eds., *The Psychology of Vision*. London: The Royal Society.

Braddick, O.J., Campbell, F.W., and Atkinson, J.

1978 Channels in vision: basic aspects. In R. Held, H.W. Leibowitz, and H.L. Teuber, eds., *Handbook of Sensory Physiology*, Vol. 8. Heidelberg: Springer.

Braunstein, M.L.

1976 *Depth Perception Through Motion*. New York: Academic Press.

1983 Contrasts between human and machine vision: should technology recapitulate phylogeny? In J. Beck, B. Hope, and A. Rosenfeld, eds., *Human and Machine Vision*. New York: Academic Press.

Brunswik, E.

1956 *Perception and the Representative Design of Psychological Experiments*, 2nd ed. Berkeley: University of California Press.

Buffart, H., Leeuwenberg, E.L.J., and Restle, F.

1981 Coding theory of visual pattern completion. *Journal of Experimental Psychology: Human Perception and Performance* 7:241–274.

Butler, D.L.

1982 Predicting the perception of three-dimensional objects from the geometrical information in drawings. *Journal of Experimental Psychology: Human Perception and Performance* 8:674–692.

Campbell, A.G., Hartwell, R., and Hood, D.

1978 Lightness constancy at the level of the frog's optic nerve fiber. *Proceedings of the Eastern Psychological Association* 49:47 (Abstract).

Campbell, F.W., and Robson, J.G.

1964 Application of Fourier analysis to the modulation response of the eye. *Journal of the Optical Society of America* 54:518A (Abstract).

Cavanaugh, P.

1984 Image transforms in the visual system. In P.C. Dodwell and T. Caelli, eds., *Figural Synthesis*. Hillsdale, N.J.: Lawrence Erlbaum Associates.

Cooper, L.A.

1976 Demonstration of a mental analog of an external rotation. *Perception and Psychophysics* 19:296–302.

1984 Strategic Factors in Complex Spatial Problem Solving. Invited paper presented at the annual meeting of the Midwestern Psychological Association, Chicago, Illinois.

Cutting, J.E.

1983 Four assumptions about invariance in perception. *Journal of Experimental Psychology: Human Perception and Performance* 9:310–317.

Cutting, J.E., and Millard, R.T.

1984 Three gradients and the perception of flat and curved surfaces. *Journal of Experimental Psychology: General* 113:198–216.

Dallenbach, K.M.

1951 A puzzle picture with a new principle of concealment. *American Journal of Psychology* 54:431–433.

Descartes, R.

1650/ 1931 Les passions de l'ame. In E.S. Haldane and G.R.T. Ross, trans., *The Philosophical Works of Descartes*. Cambridge, England: University Press.

DeValois, R., and Jacobs, G.

1968 Primate color vision. *Science* 162:533–540.

DeValois, R.L., Albrecht, D.G., and Thorell, L.G.

1976 Spatial tuning of LGN and cortical cells in monkey visual systems. Pp. 60–63 in H. Spekreijse and H. van der Tweel, eds., *Spatial Contrast*. Amsterdam: North-Holland.

Donchin, E., Ritter, W., and McCallum, W.C.

1978 Cognitive psychophysiology: the endogenous components of the ERP. Pp. 349–412 in E. Calloway, P. Tueting and S.H. Keslow, eds., *Event-Related Potentials in Man*. New York: Academic Press.

Foster, D.H.

1984 Local and global computational factors in visual pattern recognition. In P.C. Dodwell and T. Caelli, eds., *Figural Synthesis*. Hillsdale, N.J.: Lawrence Erlbaum Associates.

Galton, F.

1883 *Inquiries into Human Faculty and its Development*. London: Macmillan.

Garner, W.R., Hake, H.W., and Eriksen, C.W.

1956 Operationism and the concept of perception. *Psychological Review* 63:149–159.

Gibson, J.J.

1950 *The Perception of the Visual World*. Boston: Houghton Mifflin.

1951 What is a form? *Psychological Review* 58:403–412.

1966 *The Senses Considered as Perceptual Systems*. Boston: Houghton Mifflin.

1979 *The Ecological Approach to Visual Perception*. Boston: Houghton Mifflin.

Gillam, B.

1972 Perceived common rotary motion of ambiguous stimuli as a criterion for perceptual grouping. *Perception and Psychophysics* 11:99–101.

Ginsburg, A.

1971 Psychological Correlates of a Model of the Human Visual System. Master's thesis, Air Force Institute of Technology, Dayton, Ohio.

Gogel, W.C.

1984 The role of perceptual interrelations in figural synthesis. In P.C. Dodwell and T. Caelli, eds., *Figural Synthesis*. Hillsdale, N.J.: Lawrence Erlbaum Associates.

Graham, C.H.

1963 On some aspects of real and apparent visual movement. *Journal of the Optical Society of America* 53:1019–1025.

Graham, N.

1981 Psychophysics of spatial-frequency channels. In M. Kubovy and J. Pomerantz, eds., *Perceptual Organization*. Hillsdale, N.J.: Lawrence Erlbaum Associates.

Graham, N., and Nachmias, J.

1971 Detection of grating patterns containing two spatial frequencies: a comparison of single-channel and multiple-channel models. *Vision Research* 11:251–259.

Green, B.

1961 Figure coherence in the kinetic depth effect. *Journal of Experimental Psychology* 62:272–282.

Gregory, R.L.

1970 *The Intelligent Eye*. London: Weidenfeld.

Gross, C.G., and Mishkin, M.

1977 The neural basis of stimulus equivalence across retinal translation. Pp. 109–122 in S. Harnad et al., eds., *Lateralization in the Nervous System* New York: Academic Press.

Haber, R.N.

1983 Stimulus information and processing mechanisms in visual space perception. In J. Beck, B. Hope, and A. Rosenfeld, eds., *Human and Machine Vision*. New York: Academic Press.

Haber, R.N., and Wilkinson, L.

1982 The perceptual components of computer graphic displays. *Computer Graphics and Applications* 2(3):23–25.

Harris, C.S., ed.

1980 *Visual Coding and Adaptability*. Hillsdale, N.J.: Lawrence Erlbaum Associates.

Hartline, H.K.

1949 Inhibition of activity of visual receptors by illuminating nearby retinal elements in the *Limulus* eye. *Federation Proceedings* 8:69.

Hebb, D.O.

1949 *The Organization of Behavior*. New York: John Wiley and Sons.

Helmholtz, H.L.F., von

1866/
1911 *Treatise on Physiological Optics*. Vols. ii and iii (translated from the 3rd German edition, 1909–1911). J.P.C. Southall, ed. and trans. Rochester, N.Y.: Optical Society of America.

Hering, E.

1878/
1964 *Outlines of a Theory of the Light Sense* (originally published in 1878). L. Hurvich and D. Jameson, trans. Cambridge: Harvard University Press.

Hobbes, T.

1651/
1948 *Human Nature* (originally published in 1651). In W. Dennis, ed., *Readings in the History of Psychology*. New York: Appleton-Century-Crofts.

Hochberg, J.

1956 Perception: toward the recovery of a definition. *Psychological Review* 63:400–405.

1962 The psychophysics of pictoral perception. *Audio-Visual Communication Review* 10:22–54.

1968 In the mind's eye. In R.N. Haber, ed., *Contemporary Theory and Research in Visual Perception*. New York: Appleton-Century-Crofts.

1978a Motion Pictures of Mental Structures. Presidential address to the Eastern Psychological Association. Washington, D.C., April.

1978b *Perception*. Englewood Cliffs, N.J.: Prentice-Hall.

1981 Levels of perceptual organization. In M. Kubovy and J. Pomerantz, eds., *Perceptual Organization*. Hillsdale, N.J.: Lawrence Erlbaum Associates.

1982 How big is a stimulus? In J. Beck, ed., *Organization and Representation in Perception*. Hillsdale, N.J.: Lawrence Erlbaum Associates.

1984a Form perception: experience and explanations. In P.C. Dodwell and T. Caelli, eds., *Figural Synthesis*. Hillsdale, N.J.: Lawrence Erlbaum Associates.

1984b Visual Worlds in Collision: Invariances and Premises, Theories versus Facts. Presidential address, Division of Experimental Psychology, annual meeting of the American Psychological Association, Toronto.

Hochberg, J., and Brooks, V.

1960 The psychophysics of form: reversible-perspective drawings of spatial objects. *American Journal of Psychology* 73:337–354.

Hochberg, J., and McAlister, E.

1953 A quantitative approach to figural "goodness." *Journal of Experimental Psychology* 46:361–364.

Hochberg, J., and Spiron, J.

1985 The Ames window: unveridical "direct perception" and not perceptual inference? *Proceedings and Abstracts of the Annual Meeting of the Eastern Psychological Association* 56:38.

Hochberg, J., Amira, L., and Peterson, M.

1984 Extensions of the Schwartz/Sperling phenomenon: invariance under transformation fails in the perception of objects' moving pictures. *Proceedings and Abstracts of the Annual Meeting of the Eastern Psychological Association* 55:17 (Abstract).

von Hornbostel, E.M.

1922 Über optische inverson. *Psychologische Forschung*, 1:130–156.

Hubel, D.H., and Wiesel, T.N.

1962 Receptive fields, binocular interaction, and functional architecture in the cat's visual cortex. *Journal of Physiology* 160:106–154.

1968 Receptive fields and functional architecture of the monkey cortex. *Journal of Physiology* 195:215–243.

Hurvich, L., and Jameson, D.

1957 An opponent-process theory of color vision. *Psychological Review* 64:384–404.

1974 Opponent processes as a model of neural organization. *American Psychologist* 29:88–102.

Johansson, G.

1977 Spatial constancy and motion in visual perception. In W. Epstein, ed., *Stability and Constancy in Visual Perception*. New York: Wiley and Sons.

1980 About Perspective Transformations and the Theory of Visual Space Perception. Uppsala Psychological Reports, No. 278. Department of Psychology, University of Uppsala, Sweden.

Kabrisky, M., Tallman, T., Day, C.H., and Radoy, C.M.
1970 A theory of pattern perception based on human physiology. In A.T. Welford and L. Houssiadas, eds., *Contemporary Problems in Perception.* London: NATO Advanced Study Institute, Taylor and Francis.

Kaufman, L.
1974 *Sight and Mind.* New York: Oxford University Press.

Kaufman, L., and Williamson, S.J.
1982 Magnetic location of cortical activity. *Annals of the New York Academy of Science* 388:197–213.

Kelly, D.H.
1961 Visual responses to time-dependent stimuli. I. *Journal of the Optical Society of America* 51:422–429.

Kepler, J.
1611 Dioptrice. In W. von Dyk and M. Caspar, eds., *Gesammelte Werke* 4:1937–1963. Augsburg, Germany: Frank.

Koffka, K.
1935 *Principles of Gestalt Psychology.* New York: Harcourt, Brace.

Kohler, W.
1929 *Gestalt Psychology.* New York: Liveright.

Kolers, P.
1972 *Aspects of Motion Perception.* New York: Pergamon.

Kopfermann, H.
1935 Psychologische Untersuchungen über die Wirkung zweidimensionaler Darstellungen körperlicher Gebilde. *Psychologische Forschung* 13:293–364.

Korte, A.
1915 Kinematoskopische Untersuchungen. *Zeitschrift für Psychologie* 72:194–296.

Kosslyn, S.M.
1980 *Image and Mind.* Cambridge: Harvard University Press.

Leeuwenberg, E.L.J.
1971 A perceptual coding language for visual and auditory patterns. *American Journal of Psychology* 84:307–349.

Mach, E.
1886/ 1959 *The Analysis of Sensations and the Relation of the Physical to the Psychical* (translated by S. Waterlow from the 5th German edition, 1886). New York: Dover.

Marr, D.
1982 *Vision.* San Francisco: Freeman.

Marr, D., and Poggio, T.
1979 A computational theory of human stereo vision. *Proceedings of the Royal Society of London*, b204, 302–328.

McConkie, G.W., and Rayner, K.
1975 The span of effective stimulus during a fixation in reading. *Perception and Psychophysics* 17:578–586.

Metzger, W.
1934 Tiefenerscheinungen in optischen Bewegungsfeldern. *Psychologische Forschung* 20:195–260.

Mill, J.
1965 Analysis of the phenomena of the human mind. In R.J. Herrnstein and E.G. Boring, eds., *A Source Book in the History of Psychology.* Cambridge, Mass.: Harvard University Press.

visual cortex. *Journal of Physiology* 283:101–120.

Mueller, J.
1838/ 1965 *Handbuch der Physiologie des Menschen*, bks. V and VI. Coblenz, 1838 and 1840. Translated in 1848 by W. Baly and excerpted in R.J. Herrnstein and E.G. Boring, eds., *A Source Book in the History of Psychology*. Cambridge: Harvard University Press.

Newton, I.
1672/ 1948 A new theory of light and colors. *Philosophical Transactions of the Royal Society*. Reprinted in W. Dennis, ed., *Readings in the History of Psychology*. New York: Appleton-Century-Crofts.

Oately, K.
1978 *Perceptions and Representations*. New York: Free Press.

Pantle, A., and Sekuler, R.
1968 Size-detecting mechanism in human vision. *Science* 162:1146–1148.

Penrose, L., and Penrose, R.
1958 Impossible objects: a special type of visual illusion. *British Journal of Psychology* 49:31–33.

Perrett, D.I., Rolls, E.T., and Caan, W.
1982 Visual neurones responsive to faces in the monkey inferotemporal cortex. *Experimental Brain Research* 47:329–342.

Peterson, M.A., and Hochberg, J.
1983 Opposed-set measurement procedure: a quantitative analysis of the role of local cues and intention in form perception. *Journal of Experimental Psychology: Human Perception and Performance* 9:183–193.

Rashevsky, N.
1948 *Mathematical Biophysics*. Chicago: University of Chicago Press.

Ratliff, F.
1965 *Mach Bands: Quantitative Studies on Neural Networks in the Retina*. San Francisco: Holden-Day.

Reite, M., and Zimmerman, J.
1978 Magnetic phenomena of the central nervous system. *Annual Review of Biophysics and Bioengineering* 7:167–188.

Restle, F.
1979 Coding theory of the perception of motion configurations. *Psychological Review* 86:1–24.

Roberts, L.G.
1965 Machine perception of three-dimensional solids. In J.T. Tippett et al., eds., *Optical and Electro-Optical Information Processing*. Cambridge: MIT Press.

Rock, I.
1977 In defense of unconscious inference. In W. Epstein, ed., *Stability and Constancy in Visual Perception*. New York: John Wiley and Sons.
1983 *The Logic of Perception*. Cambridge: MIT Press.

Rosenblatt, F.
1962 *Principles of Neurodynamics*. New York: Spartan Books.

Schade, O.H.
1956 Optical and photoelectric analog of the eye. *Journal of the Optical Society of America* 46:721–739.

Schuck, J., and Leahy, W.R.
1966 A comparison of verbal and non-verbal reports of fragmenting visual images. *Perception and Psychophysics* 1:191–192.

Schwartz, B.J., and Sperling, G.
1983 Luminance controls the perceived 3-D structure of dynamic 2-D displays. *Bulletin of the Psychonomic Society* 21(6):456–458.

Selfridge, O.G.
1959 Pandemonium: a paradigm for learning. In *The Mechanization of Thought Processes.* London: H.M. Stationery Office.

Shepard, R.N.
1981 Psychophysical complementarity. In M. Kubovy and J.R. Pomerantz, eds., *Perceptual Organization*. Hillsdale, N.J.: Lawrence Erlbaum Associates.

Shepard, R.N., and Cooper, L.
1982 *Mental Images and Their Transformations*. Cambridge: MIT-Bradford Books.

Shepard, R.N., and Metzler, J.
1971 Mental rotation of three-dimensional objects. *Science* 171:701–703.

Sperling, G.
1970 Binocular vision: a physical and a neural theory. *American Journal of Psychology* 83:461–534.

Stevens, K.A.
1983 The visual interpretation of surface contours. *Artificial Intelligence* 17:47–73.

Sutton, S., Braden, M., Zubin, J., and John, E.R.
1965 Evoked potential correlates of stimulus uncertainty. *Science* 150:1187–1188.

Svaetichin, G.
1956 Spectral response curves from single cones. *Acta Physiologica Scandinavica* 39(Suppl. 134):17–46.

Todd, J.
1982 Visual information about rigid and nonrigid motion: a geometric analysis. *Journal of Experimental Psychology: Human Perception and Performance* 8:238–252.

Todd, J.T., and Mingola, E.
1983 Perception of surface curvature and direction of illumination from patterns of shading. *Journal of Experimental Psychology: Human Perception and Performance* 9:583–595.

Tolman, E.C.
1938 Schematic "sowbug" and discrimination learning. *Psychological Bulletin* 35:524.

Ullman, S.
1979 *The Interpretation of Visual Motion*. Cambridge: MIT Press.

Wallach, H.
1948 Brightness constancy and the nature of achromatic colors. *Journal of Experimental Psychology* 38:310–324.

Wallach, H., and O'Connell, D.N.
1953 The kinetic depth effect. *Journal of Experimental Psychology* 38:310–324.

Watson, J.B.
1913 Psychology as the behaviorist views it. *Psychological Review* 20:158–177.

Wheatstone, C.
1839 On some remarkable and hitherto unobserved phenomena of binocular vision. Part 2. *Philosophical Magazine* 4:504–523.

White, B., and Mueser, G.
1960 Accuracy of reconstructing the arrangement of elements generating kinetic depth displays. *Journal of Experimental Psychology* 60:1–11.

Wohlegemuth, A.
1911 On the aftereffect of seen movement. *British Journal of Psychology* Monograph Supplement, 1.

Woodworth, R.S.
1938 *Experimental Psychology*. New York: Holt, Rinehart and Winston.